The American Economic Review

ARTICLES

87

DECEMBER 1985

THE AMERICAN ECONOMIC ASSOCIATION

•Printed at Banta Company, Menasha, Wisconsin. The publication number is ISSN 0002-8282.

•*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the *Survey*, and Supplements, is published by the American Economic Association and is sent to all members six times a year: March; May; June; September; semimonthly, December.

Regular member dues (nonrefundable) for 1986, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature* are as follows:

- \$37.50 if annual income is \$30,000 or less;
- \$45.00 if annual income is above \$30,000, but no more than \$40,000;
- \$52.50 if annual income is above \$40,000;
- \$18.75 annually for registered students (three years only).

Nonmember subscriptions will be accepted only for both journals: Institutions (libraries, firms, etc.), \$105 a year; individuals, \$70.00. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add \$12.00 to the annual rates above to cover extra postage.

•Correspondence relating to the *Survey*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

•Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

•Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

Founded in 1885

Robert W. Clower

Managing Editor of the *American Economic Review*

1980–85

THE AMERICAN ECONOMIC REVIEW

- Submit manuscripts (4 copies), no longer than 50 pages, double spaced, to:
  Orley Ashenfelter, Managing Editor, *AER*; 169 Nassau Street, Princeton, NJ 08542–7067.
- Submission fee: $25 for members; $50 for nonmembers. Style guides will be provided upon request.
- No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

December 1985

VOLUME 75, NUMBER 5

Articles

Shorter Papers

The following Statement of Ownership, Management, and Circulation is provided in accordance with the requirements, as contained in 39 U.S. Code 3685. The *American Economic Review* is owned, managed, and published by the American Economic Association, a nonprofit educational organization, located at 1313 21st Avenue South, Nashville, Davidson County, Tennessee 37212. The Managing Editor is Professor Orley Ashenfelter, *American Economic Review*, 169 Nassau Street, Princeton, NJ 08542–7067. During the preceding 12 months the average number of copies printed for each issue was 27,700; the average paid circulation, 23,941; the average free distribution, 403; the average number of copies distributed, 24,344. Corresponding figures for the last issue before filing: 27,500 total number copies printed; 25,206 total paid circulation; 398 free copies distributed; 25,604 total distribution.

New Estimates of the Value of Federal Mineral Rights and Land

By MICHAEL J. BOSKIN, MARC S. ROBINSON, TERRANCE O'REILLY, AND PRAVEEN KUMAR*

The federal government owns a large fraction of the mineral rights and land in the United States. The value of these resources varies substantially over time, both with acquisitions and sales of these assets and with changes in the prices of the minerals and land involved. Whether one is interested in measures of national wealth, land management policy as part of efficient government operations, or (long-run) macroeconomic fiscal policy issues, time-series estimates of the value of federal mineral rights and land are potentially valuable information.

At one extreme, consider a country or state that owns substantial mineral rights when the price of those minerals skyrockets. The additional revenues potentially available either for use in the public sector, or to allow tax cuts to provide greater private income, may alter the course of the economy. There are historical precedents. Saudi Arabia was able virtually to abolish taxation due to revenues from the sale of mineral rights, and Alaska used oil-based revenues to provide cash grants on a per capita basis to its citizens, thereby increasing their private wealth and consumption opportunities.

The microeconomics of sensible resource allocation, within the public sector and between the public and private sector, relies on careful cost-benefit evaluations of the value of public services and the opportunity costs of providing them. These in turn are only possible with accurate information on actual *and* potential revenue sources, including the opportunity costs of purchases and sales of assets, as well as traditional flows of income into and out of the public sector. Therefore, the value of federal mineral rights and land is potentially an important piece of information for a host of public policy questions.

This paper provides, we believe for the first time, estimates of the value of federal mineral rights in the postwar period in the United States. It also presents a new time-series for the value of federal land and compares the results with previous estimates. We find that the values of federal mineral rights and land are enormous and fluctuate by tens of billions of dollars annually.

In Section I, we consider the valuation of federal mineral rights; in particular, the revenue which the government is able to obtain from onshore and offshore leases for the extraction of oil and natural gas. We review some recent studies of resource accounting (none of which focuses on the government sector) and present a new methodology for valuing mineral rights. We estimate the value of federal oil and gas mineral rights to be \$819 billion in 1981, a number higher than the privately held federal debt in that year.[1] Our estimates consider economically recover-

*Boskin: Professor of Economics, Stanford University, Stanford, CA 94305 and Research Associate, National Bureau of Economic Research; Robinson: Assistant Professor of Economics, University of California-Los Angeles, Los Angeles, CA 90024; O'Reilly and Kumar: graduate students, Economics Department, Stanford University. This paper is part of Boskin's larger project on more comprehensive federal government budgets. We are indebted to Dennis Epple, an anonymous referee, and seminar participants at Stanford University and the National Bureau of Economic Research for helpful comments. We thank the Center for Economic Policy Research at Stanford University and the National Bureau of Economic Research for financial support.

[1]The total national debt in 1981 was \$1,004 billion. Of this, \$210 billion was held by government agencies and the Federal Reserve, leaving \$794 billion held by private individuals and institutions.

able undiscovered reserves,[2] and therefore would substantially exceed estimates that used the methods of previous studies.

Section II is devoted to the valuation of federal land. We review the work of Raymond Goldsmith (1962), Grace Milgram (1973), and Robert Eisner and Paul Pieper (1984) on this subject, and we present new updated estimates of the value of federal land, taking into account the changing composition of federal land. Our estimate for 1981 is $175 billion, composed of $112 billion for urban land and $63 billion for rural land.

Section III provides a summary and agenda for research, and the Appendix provides details of Milgram's methodology for estimating government land value and our extension of her estimates.

I. The Value of Federal Mineral Rights

A. *Previous Studies of Resource Accounting*

Although no previous study attempts to value federal mineral rights specifically, a number of recent studies of income accounting for exhaustible resources, such as oil and natural gas, support the inclusion of estimates of the underground reserves of these resources in measures of national income and wealth.[3] The current accounting practice is to exclude such estimates. The National Income and Product Accounts of the Bureau of Economic Analysis consider only production of mineral resources, ignoring the level of reserves. The studies argue for the inclusion of the net value of *proven* reserves in estimates of national wealth.[4] This value could change through discovery, depletion, and changes in the price of the resource.

The perpetual inventory method provides a means of building on an estimate of the value of proven minerals for a particular year. Determining the annual changes in value is fairly straightforward; the major difficulty arises in determining a base year estimate for the value of proven resources. In an important study, J. Steven Landefeld and James Hines discuss three methods for estimation: the present value method, the land price method, and the net price method.

The present value method requires forecasting prices, operating costs, production and interest rates over the life of the field after its discovery. The present value of the stream of net revenues is determined. John Soladay extends this method by attempting to take into account the upward revisions in estimates of reserves that typically occur after the discovery; the total quantity produced from a field is greater than the initial discovery. Several ways of guessing future trends in net revenues have been used. Soladay extrapolates future net revenues based on the weighted average of net revenues over the period 1948–74 and chooses a particular interest rate. Landefeld and Hines report results for three arbitrary choices of growth rate in net revenue and interest rates. The Securities and Exchange Commission (SEC)[5] proposed that companies be required to assume no growth in net revenue and a 10 percent discount rate.

The net price method assumes that net revenues increase at the rate of interest. According to economic theory, this is necessary for equilibrium if the cost of exploration and extraction is the same for all of the exhaustible resource.[6] The advantage of this method is that it does not require any assumptions regarding the time path of production, since any pattern has the same present value.

[2] Those resources estimated to be recoverable and profitable to extract at current prices and technology.

[3] See the U.S. Securities and Exchange Commission (1979), UN Economic and Social Council (1979; 1980), Financial Accounting Standards Board (1980), John Soladay (1980), and J. Steven Landefeld and James R. Hines (1982).

[4] They also favor the incorporation of this value in accounting measures of firm wealth.

[5] Securities and Exchange Commission (p. 503) quoted in Landefeld and Hines, p. 150.

[6] If the resource has an increasing cost of extraction, net price should increase at less than the rate of interest in equilibrium. Landefeld and Hines neglect to note this in their defense of the net price method.

The third method discussed by Landefeld and Hines, the land price method, assumes that the entire value of the proven reserves is paid to the landowner in the form of bonus payments and royalties.[7] They also assume that royalties are a constant fraction of the net value of the resource, so that annual data on bonuses can be used to estimate the net value of the oil and gas prospects leased that year. One problem with this method is that the value of oil and gas prospects leased in a particular year bears no particular relationship, even in expected value terms, to new proven reserves in that year, because of decision and drilling lags. This is therefore fundamentally different from either the present value or net price methods. A second difficulty lies in Landefeld and Hines' estimates, since they use 12.5 percent as the fraction of net price which is paid in the form of royalties. Since royalties are at least 12.5 percent of the *gross* price of oil and gas, this significantly understates the importance of royalties, and their estimates for the value of oil and gas prospects leased are too low. This is confirmed by noting that the estimates using the land price are much smaller than their estimates using other methods. (See Landefeld-Hines, p. 159.)

Each of these methods is inadequate for creating *government* wealth and capital formation accounts. The most important problem is the neglect of the value of economically recoverable undiscovered reserves. This neglect, as argued below, causes an understatement of wealth and capital gains and a misstatement of government investment.

Earlier studies argued against including economically recoverable undiscovered resources in either national or firm accounts on the grounds that estimates were too uncertain. This problem is much less severe for the United States as a whole than for the individual firm, since the sample of prospects is far larger and, therefore, the distribution is tighter (the coefficient of variation is smaller). While the range of estimates of undiscovered resources may be wide, there is no a priori reason for believing an estimate to be biased. By contrast, assuming that undiscovered resources have no value is surely biased and, therefore, estimates of wealth and income will be biased.

B. *Methodology for Valuing Mineral Rights*

When the government leases the mineral rights in a particular area—rights essentially to as yet undiscovered resources—it has reduced its mineral wealth by transferring claims to part of it to the private sector. In return, the government receives some payment immediately in the form of a bonus, with the rest of the payments deferred as royalties or rental payments. Bonuses are cash payments that are not conditional on the existence or size of the resource, and are typically the variable subject to bidding. Royalty payments are fractions, usually fixed in advance, of the gross revenue of the produced output, if any. By the time reserves are proven, their only value to the government is the present value of the royalties they represent.

The base-year value to the government of federal mineral rights is the sum of three components: future royalties on proven reserves; future royalties on estimated undiscovered reserves; and future bonuses on unleased land.[8] This may be written (choosing 1981 as the base year) as follows:

$$(1) \qquad V_{1981} = PVR_p + PVR_u + PVB,$$

where PVR_p = present value of future royalties on proven reserves, both onshore and offshore; PVR_u = present value of future

[7]Payments to landowners are only relevant to calculations concerning firms, and not to calculations of national wealth.

[8]This is obviously not the total value of the minerals on federal land, nor is it even necessarily the remaining scarcity rent on the minerals (since bonuses and royalties may not capture the full rent). We ignore rental payments, that are quite small compared to bonuses and royalties, as well as additional taxes that might be generated from the production. The extra taxes should only arise from special taxes on the resource (for example, crude oil windfall profits tax) or the economic surplus captured by the producer rather than the government.

royalties on undiscovered reserves, again both onshore and offshore; and PVB = present value of bonuses on mineral leases.

To obtain the value for any future year, we take the value for the previous year, add capital gains or losses, and subtract bonus and royalty payments received. Capital gains and losses are calculated by assuming that the current price is the base from which future prices grow at the interest rate.[9] Since all three components of the base-year value are proportional to the current price, the capital gain is just the change in price times the previous year's value. Using this method, the base-year value can be projected backwards as well.

Ignoring undiscovered reserves can cause several problems in the wealth and income accounts of the government. For example, the sale of leases would be treated as an increase in government receipts and wealth rather than an asset sale, and future royalty rights would not appear in the accounts until drilling was successful. Further, capital gains and losses associated with price changes would only be counted on proven reserves. Government capital formation, defined as the change in government wealth, would be overstated, since the sale of assets in the form of possible reserves would be ignored.[10]

For all of these reasons, we believe accurate resource accounting for the government sector requires estimating a value for the undiscovered reserves on government land. This need only be done for a base year; the perpetual inventory method may then be applied to calculate other years.[11] For the base-year calculation, we use U.S. Department of the Interior estimates for the expected undiscovered reserves for onshore and offshore federal land.[12] To value the royalties on these undiscovered reserves, an assumption regarding future prices needs to be made. We choose the strong and convenient assumption that future prices are expected to increase at the rate of interest.[13] This assumption means that the time path of production is irrelevant; all production patterns yield the same present value of royalties. Such time independence is especially useful for undiscovered reserves, since there is a substantial and uncertain time until the resource will be extracted. The assumption can be justified by noting that the Long-Term Pricing Committee of the Organization of Petroleum Exporting Countries (OPEC) has recommended a 3 percent real annual increase in oil prices that, given its low cost of production, would be close to that suggested by economic theory. The average annual rise in real oil prices received by U.S. producers was 3.5 percent over the period 1950–82.[14]

This assumption implies that the value of future royalties on both undiscovered and proven reserves on federal land is the royalty rate times the quantity of reserves times the current price for the relevant resource. Thus,

$$(2)\quad PVR_p = \sum_k P_k \times \left(r^f \times R_p^{kf} + r^o \times R_p^{ko}\right),$$

where k indexes the mineral, f indexes offshore reserves, o indexes onshore reserves, R is the quantity of reserves, and r is the relevant royalty rate.[15]

[9] Capital gains are in current dollars; they occur even when the prices increase as predicted. We consider alternative price increase assumptions below.

[10] All of these distortions occur in the accounts of any landowner if only proven reserves are taken into consideration. The earlier studies, discussed in the previous section, were concerned with valuing the assets and depletion of a producer; they were not concerned as much with the landowner from whom the lease was obtained. In this paper, of course, the government is the focus; however, in the United States, the government generally is not a producer of minerals.

[11] This does not take into account changes in the estimated recoverable undiscovered reserves. These would be paper gains and losses. We examine alternative estimates of undiscovered resources below.

[12] The estimates of offshore undiscovered reserves were obtained from the *Federal Offshore Statistics* (1983). The corresponding onshore estimates were obtained by personal correspondence with D. Zimmerman of the Department of Interior. We discuss the methods used by the department below.

[13] We consider alternative assumptions later.

[14] Calculated using price data from the American Petroleum Institute (1984).

[15] The royalty rate is assumed to be 12.5 percent for onshore federal land and 16.67 percent for offshore reserves in the case of oil and natural gas. It is calculated using the ratio of minerals produced to royalty payment received from U.S. Department of the Interior, *Mineral Revenues* (1983). This source also permits the

Similarly,

$$PVR_u = \sum_k P_k \times \left(r^f \times R_u^{kf} + r^o \times R_u^{ko}\right). \tag{3}$$

The present value of future bonus payments on unleased land also needs to be included to obtain the base-year estimate for the value of federal mineral rights. To do this, we first find the present value of bonuses paid to the federal government over the period 1954–79.[16] We divide this by the present value of royalties paid over the period 1956–81. The difference in the periods covered is designed to account for discovery and production lags. We assume that the present value of future bonuses on undiscovered resources will be the same fraction of estimated future royalties on these undiscovered resources as occurred in this period.[17]

Thus, we assume,

$$\frac{PVB_{1954-79}^{1981}}{PVR_p + PVR_{1956-81}^{1981}} = \frac{PVB}{PVR_u}, \tag{4}$$

where $PVB_{1954-79}^{1981} = 1981$ present value of bonuses paid to the government on leased land from 1954 to 1979, in 1981 prices, and $PVR_{1956-81}^{1981} = 1981$ present value of royalties paid to the government from 1956 to 1981, in 1981 prices.

Since PVR_p and PVR_u are calculated as in (2) and (3) above, we only need to convert bonuses and royalties actually paid in the period to present value dollars. To do this, we assume an annual real rate of discount of 2 percent and compute

$$PVB_{1954-79}^{1981} = \sum_{\tau=1954}^{1979} B_\tau (1.02)^{(1981-\tau)} \cdot \frac{Q_{1981}}{Q_\tau}, \tag{5}$$

where B_τ = bonuses paid to the government for year τ, $Q_{1981} = GNP$ deflator for 1981, and $Q_\tau = GNP$ deflator for year τ.

Similarly,

$$PVR_{1956-81}^{1981} = \sum_{\tau=1956}^{1981} R_\tau (1+0.2)^{(1981-\tau)} \cdot \frac{Q_{1981}}{Q_\tau}, \tag{6}$$

where R_τ = royalties received by the government in year τ.

C. *Estimates of the Value of Federal Oil and Gas Mineral Rights*

Our estimates for the value of federal oil and gas mineral rights in 1981 are presented in Table 1. Two striking facts are apparent. First, the present value of bonuses from offshore mineral leases far exceeds the corresponding figure for onshore leases. The reason for this is that offshore bonuses were much greater than onshore bonuses in the period 1954 to 1979. For example, since 1971, offshore bonuses have annually exceeded \$1 billion whereas onshore bonuses did not reach \$20 million (see *Mineral Revenues*). This difference is reflected in our estimate of the present value of bonuses through equation (4).

Second, the present value of future royal-

TABLE 1—CALCULATION OF THE 1981 VALUE OF FEDERAL MINERAL RIGHTS FOR OIL AND NATURAL GAS (V_{1981}) (Billions of 1981 Dollars)

Component	Total	Onshore	Offshore
PVB	221.1	0.9[a]	220.1
PVR_p	88.4	34.5	53.9
PVR_u	509.8	111.0[b]	398.8
total: V_{1981}[c]	819.3	146.4	672.8

[a]This figure is calculated assuming that the future ratio of onshore and offshore bonuses will remain the same as in the historical period.

[b]This figure is derived assuming that the ratio of offshore undiscovered gas reserves in Alaska is the same as in the 48 states.

[c]The present value of bonuses for the period 1954–79 for offshore and onshore were \$50.9 and \$0.2 billion, respectively. The corresponding figures for the value of royalties for the period 1956–81 were \$20.81 and \$8.70 billion, respectively.

calculation of royalty rates for other minerals. If the federal government is forced to share royalties with the states, the value of the mineral rights developed later would be divided with them.

[16]See *Mineral Revenues* (Tables 10 and 13).

[17]Changes in development lags would not substantially alter the results. The assumption that the ratio of bonuses to royalties remained constant seems reasonable given that real prices are expected to increase; it is not founded in any particular model.

TABLE 2—VALUE OF FEDERAL OIL AND NATURAL GAS RIGHTS AND CHANGES IN VALUE, 1954–82
(Billions of Current Dollars)

Year	Value			Change in Value
	Total[a]	Oil	Gas	
1954	80.6	62.3	18.3	–
1955	80.8	62.0	18.8	0.2
1956	81.9	62.4	19.5	1.1
1957	89.4	69.0	20.4	7.5
1958	88.7	67.2	21.5	−1.7
1959	87.8	64.6	23.2	−0.9
1960	90.0	63.9	25.1	2.2
1961	91.1	64.0	27.1	1.1
1962	91.3	63.7	27.6	0.2
1963	91.5	63.4	28.1	0.2
1964	90.4	63.0	27.4	−1.1
1965	90.1	62.4	27.7	−0.3
1966	90.2	62.5	27.7	0.1
1967	90.9	62.8	28.1	0.7
1968	90.4	62.1	28.3	−0.5
1969	93.6	64.9	28.7	3.2
1970	94.9	65.9	28.7	3.2
1971	100.6	69.8	30.8	5.7
1972	101.5	67.9	30.6	0.9
1973	109.9	75.4	34.5	8.4
1974	176.0	129.2	46.8	66.1
1975	210.8	142.9	67.9	34.8
1976	238.0	150.5	87.5	27.2
1977	273.9	155.8	118.1	35.9
1978	295.7	161.7	134.0	21.8
1979	389.5	222.8	166.7	93.8
1980	598.3	376.4	221.9	208.8
1981	819.3	547.1	272.2	221.0
1982	817.2	486.8	330.4	−2.1

[a]We have assumed that the present values of gas and oil bonuses are proportional to the present values of gas and oil royalties on undiscovered reserves; i.e, $PVB^{gas}/PVB^{oil} = PVR_u^{gas}/PVR_u^{oil}$.

ties from economically recoverable undiscovered reserves similarly dominates the corresponding figure for proven reserves. The explanation for this is straightforward: estimates of undiscovered resources are much larger than currently proven reserves.

The detailed time-series of the total value of federal oil and gas mineral rights from 1954 to 1982, presented in Table 2, are extremely interesting. The aggregate series began a very rapid growth in 1974, and jumped again in 1979–80. The current value of over $800 billion is the single largest asset in the complete balance sheet of the federal government. It is substantially larger than the value of federal land. In fact, it is approximately the combined value of *all* federal tangible assets or *all* federal financial assets (see Eisner and Pieper). Prior to 1974, the total series was quite stable in nominal dollars, and therefore it exhibited a slight downward decline in real terms. While the dollar value has increased sharply recently, even prior to the increases in energy prices the value for oil and gas were substantial. For example, in 1971 the value was $100 billion (in 1971 dollars) which was much more than the value of federal land. It also was twice as large as the value of federal government gold holdings.

The relative value of oil and gas in the total has changed somewhat over the period. While oil is still the largest component, the share of oil has fallen from over three-fourths in the late 1950's, to two-thirds or less in the last few years. While oil typically receives more attention than natural gas, these figures reveal the importance of natural gas.

TABLE 3—COMPONENTS OF THE CHANGE IN VALUE OF FEDERAL OIL AND GAS RIGHTS, 1954–82
(Billions of Current Dollars)

		Components		
Year	Change in Value	Revaluations	Bonuses	Royalties
1954	–	–	0.1	0.0
1955	0.2	0.3	0.1	0.1
1956	1.1	1.2	0.0	0.1
1957	7.5	7.6	0.0	0.1
1958	−1.7	−1.6	0.0	0.1
1959	−0.9	−0.7	0.1	0.1
1960	2.2	2.6	0.3	0.1
1961	1.1	1.2	0.0	0.1
1962	0.2	0.8	0.5	0.1
1963	0.2	0.4	0.0	0.2
1964	−1.1	−0.8	0.1	0.2
1965	−0.3	−0.1	0.0	0.2
1966	0.1	0.5	0.2	0.2
1967	0.7	1.4	0.5	0.2
1968	−0.5	1.1	1.3	0.3
1969	3.2	3.6	0.1	0.3
1970	1.3	2.6	0.9	0.4
1971	5.7	6.2	0.1	0.4
1972	0.9	3.7	2.3	0.5
1973	8.4	12.0	3.1	0.5
1974	66.1	71.8	5.0	0.7
1975	34.8	36.7	1.1	0.8
1976	27.2	30.3	2.2	0.9
1977	35.9	38.7	1.6	1.2
1978	21.8	25.0	1.8	1.4
1979	93.8	100.8	5.1	1.9
1980	208.8	215.8	4.2	2.8
1981	221.0	231.9	6.7	4.2
1982	−2.1	6.8	4.1	4.8

Also included in Table 2 is the change in value from year to year. These changes tended to be small until the total value became large subsequent to the substantial increases in energy prices in 1973–74. The change in the value of these mineral rights in many years in the 1970's and early 1980's exceeded the nominal federal government budget deficit (see Boskin, 1982).

Table 3 breaks the change in the value of federal oil and natural gas mineral rights through time into three components: revaluation; bonuses; and royalties. While bonuses and royalties became large in the early 1970's, they are still relatively minor compared with the enormous revaluations of this period. The bulk of the change in the value in most years is the revaluation of the rights. The revaluations largely reflect the energy price shocks, but once the total value of oil and gas becomes large, even small price changes can lead to large revaluations. It should be stressed that revaluations would occur even if the price followed the assumed pattern of growing at the interest rate. In recent years, these revaluations are substantial relative to the capital gains on assets held by the household sector of the United States (see Eisner, 1980).

The figures in Tables 2 and 3 reveal how important the value of federal government mineral rights can be to measures of national wealth, to measures of changes in that wealth, to mineral leasing policy, and to sensible government budget reporting and policy. The total value of these mineral rights is enormous ($819 billion). To place this in perspective, in 1981, this value exceeded the value of the privately held national debt ($794 billion). Obviously, the value of other minerals

TABLE 4—PROVEN AND UNPROVEN OIL AND GAS RESERVES (1981)
(Oil in Billion Barrels/Gas in Trillion Cubic Feet)

	95% Confidence Level[a]		Mean		5% Confidence Level[a]	
	Oil	Gas	Oil	Gas	Oil	Gas
Offshore[b]						
Proven	3.8	41.6	3.8	41.6	3.8	41.6
Undiscovered, Economically recoverable	15.5	82.9	25.5	140.2	39.9	216.25
Onshore[c]						
Proven	4.6	24.9	4.6	24.9	4.6	24.9
Undiscovered, Economically recoverable	24.3	61.3	41.61	112.1	66.6	170.1

[a] Ratios to mean values for onshore reserves are assumed to be the same on federal land as total land. Also, it was assumed that the ratio of unproven offshore and onshore reserves in Alaska is the same as in the 48 states.

[b] Offshore figures are from Tables 53 and 54 of *Federal Offshore Statistics*, December 1983.

[c] Onshore figures were published only through 1980. The updated *USGS* mean figures for 1981, which are somewhat *lower* than the published 1980 figures, were kindly provided by Dale Zimmerman, U.S. Department of the Interior.

would add to this total. Clearly, ignoring the value of resources in government budgets and in national income and wealth accounts can be quite misleading.

The critical estimates of undiscovered economically recoverable resources of conventional oil and gas on federal land in 1981 were made by the U.S. Department of the Interior.[18] These are presented in Table 4. Estimating proven reserves in a field where hydrocarbons have been discovered is difficult and results in frequent revisions; the task of estimating undiscovered recoverable resources is much more complex, as the number of dry holes attests. Using a point estimate, however well-founded in expert geological opinion, perhaps suggests more certainty than actually exists. We have therefore calculated the value of federal oil and gas mineral rights for 1981 using the high (5 percent) and low (95 percent) bounds calculated by the U.S. Geological Survey (*USGS*).[19] The high estimate is $1134.9 billion, while the low estimate is $582.1 billion. While the range of these estimates is clearly large, even the low estimate shows that federal oil and gas mineral rights have considerable value. Independent studies by other groups have come up with different ranges for the quantities of undiscovered re-

[18] The U.S. Geological Survey (*USGS*) divided the U.S. onshore and offshore areas into 137 provinces. Individual appraisals were made for each of the provinces using geological data and exploration histories as the basis for separate subjective assessments by six geologists. The subjective assessments of high, modal, and low probabilities for undiscovered recoverable resources of oil and gas are averaged, then aggregated probabilistically across provinces to obtain estimates for the entire United States. A more complete description of the methodology used by the *USGS* is given in their Circular 860 (1981).

[19] The procedure for deriving the high and the low figures is as follows. The mean PVR_u figures for gas and oil are multiplied by the ratio of the corresponding 5th and 95th fractile figures (separating out the offshore and onshore components) from *USGS* and the PVR_u^{low} and PVR_u^{high} are obtained. Following the methodology for calculation of PVB, PVB^{low}, and PVB^{high} are obtained and the appropriate total values are derived according to equation (1).

coverable oil and gas in the United States, but, at least since 1975, there has been a growing consensus, with overlapping ranges and point estimates for both oil and gas approximately within the range of the *USGS* estimates.[20]

Two additional points should be made about undiscovered resources. First, they do not stay undiscovered forever. Annual additions to proven offshore reserves were about 14 percent of the total stock of proven reserves over the period 1977–81 for both oil and gas (see *Federal Offshore Statistics*). Second, estimates of recoverable undiscovered resources must properly depend upon prices and technology. Technological advances or real price increases should lead to upward revisions in the estimates of undiscovered recoverable resources.

We also tested the sensitivity of our results to alternative assumptions about the rate of price increases. Two considerations suggest that price may not grow at the rate of interest. If costs are nontrivial (as on the outer continental shelf), the scarcity rent, rather than price, should grow at the rate of interest. If the relevant costs at the margin are those of major OPEC producers, say Saudi Arabia, costs are indeed trivial, but if monopoly power is continuously exercised, marginal revenue, not price, will grow at the rate of interest. While we believe that for long-run considerations, the competitive, trivial costs result may be most appropriate, we present below estimates of the relative value of mineral rights if price grows less rapidly than the interest rate.

The net present value of proven (R_t) and undiscovered (U_t) reserves at t are given by

$$(7)\quad NPVR_t = R_t P_t \sum_{i=0}^{\infty} (1-y)(ydq)^i = R_t P_t \frac{(1-y)}{1-ydq},$$

$$(8)\quad NPVU_t = U_t P_t \sum_{j=0}^{\infty} (1-c)(cdq)^j \cdot \sum_{i=0}^{\infty} (1-y)(ydq)^i = \frac{U_t P_t (1-y)(1-c)}{(1-cdq)(1-ydq)},$$

where R_t = proved reserves at t; U_t = undiscovered (economically recoverable at current cost) reserves at t; $(1-y)$ = fraction of R_t produced during t (assumed constant); $(1-c)$ = fraction of U_t converted to R_t during t (assumed constant); P_t = real price per unit at date t; t = valuation date; d = discount factor $=1/1+r$, where r is real interest rate; and q = price growth factor.

Some sample calculations indicate the potential difference under alternative price path assumptions when price grows at less than the interest rate. Calculations assume $r=0.03$; $y=0.10$; $c=0.03$. For example, the present value of royalties with a price increase two-thirds of the assumed real interest rate is \$470 billion in 1981 as opposed to our estimate of \$598 billion. The present value of bonuses likewise would be somewhat smaller. Our point, however, that the value of federal mineral rights is large relative to other federal assets and liabilities remains unaltered. Further, virtually all of our other assumptions tend to bias the estimated value downward. (See Table 5.)

II. The Value of Federal Land

A. *Previous Studies of the Value of Federal Land*

A time-series for the value of federal land in the postwar period (1945–81) is provided by the estimates of Goldsmith (1962), and the follow-up studies by Milgram and by Eisner and Pieper. These studies, as well as the current one, demonstrate how successive refinements of basic data often hang from a very slender thread.

Goldsmith bases his postwar time-series on an estimate of the value of government land on December 31, 1946 in the study by

[20] See *USGS* for a survey of other studies.

TABLE 5–RATIO OF VALUE FOR $Q < R$

Assumed Rate of Real Price Increase[a]	$NPVR$	$NPVU$
1.0	0.851	0.552
1.5	0.884	0.680
2.0	0.920	0.761
2.5	0.958	0.864
3.0	1.000	1.000

[a]Percent per annum.

J. E. Reeve et al. (1950). The land value estimate of the Reeve et al. study has two main components: military and nonmilitary government land. These estimates are as follows: land, nonmilitary: \$4.93 billion; and land, military: (a) market value, \$1.07 billion; (b) replacement value, \$2.13 billion.

The estimate for nonmilitary land is based upon its original acquisition costs, with (casually justified) adjustments to reflect 1946 market values. Separate adjustments were made for each of five categories of land. For most categories, Reeve et al. do not cite any source as the basis for the adjustment factors employed.

For military land, Goldsmith chooses Reeve et. al.'s replacement value figure, which is an estimate of the cost to the military to replace its holdings with comparable land. For nonmilitary government land, Goldsmith's extension (1962) of the Reeve et al. estimates from 1946 through 1958 is based upon two price indices: forest land: Reeve et al.'s value for forest land in 1946 multiplied by "index of stumpage prices in national forests" (p. 380), and other civilian land: value for 1946 (Reeve et al.'s total minus forest land) multiplied by "index of grazing land prices in western states" (p. 380). Goldsmith describes his extension for military land as a "rough estimate" (p. 380).

Milgram estimates the value of government land over the period 1952 to 1968. Her series is based upon Goldsmith's estimate for 1956. Values for 1952–55 and 1957–68 are extrapolated using the following methodology. Increments to the stock of government land are derived from the change in the General Services Administration's (GSA's) annual estimates of the value of government land, from the change in acreage of government urban land and rural land, and from price series developed by Milgram for rural, nonmetropolitan, and urban land.

Milgram derives an independent estimate of the value of rural government land in 1956, using rural government acreage and a price for rural public land. The price is based primarily upon an estimate of the market value of the public domain managed by the Bureau of Land Management (BLM). The value of urban government land in 1956 is determined as the difference of Goldsmith's 1956 estimate for total federal land, and her rural land value estimate.

Increments to the rural land stock are determined by the value of the change in government owned rural acreage, using the rural public land series. Increments to the urban land stock are estimated by subtracting the incremental rural estimates from the change in the annual GSA estimate of the value of total government land. The yearly stock estimates are computed by adding these increments to the previous year's stock, after the rural component for the prior year has been adjusted by a price index for nonmetropolitan land, and the urban component for the prior year has been adjusted by an urban price index.

Eisner and Pieper estimate the value of federal land in 1980 as \$119.5 billion. They use Milgram's 1968 figure as a base and assume that annual net investment in land is zero. They infer the change in the market

TABLE 6—VALUE OF FEDERAL LAND IN BILLIONS OF (CURRENT) DOLLARS

Year	Value	Source[a]
1946	7.0	G
1951	13.4	G
1956	13.4	G, M
1961	20.6	M
1966	29.5	M
1971	37.6	E-P
1976	73.3	E-P
1980	119.5	E-P

[a]G = Goldsmith; M = Milgram; and E-P = Eisner and Pieper.

TABLE 7—VALUE OF FEDERAL LAND IN BILLIONS OF (CURRENT) DOLLARS

	New Extended Estimates			Eisner-Pieper Estimates
	Total	Urban	Rural	
1969	37.3	21.9	15.5	34.9
1970	44.8	29.0	15.8	36.4
1971	53.8	36.2	17.6	37.6
1972	63.4	44.3	19.1	42.6
1973	72.8	53.1	19.7	50.4
1974	76.5	49.9	26.6	57.3
1975	80.4	52.9	27.5	63.6
1976	90.5	57.9	32.6	73.3
1977	105.5	67.6	37.9	82.2
1978	120.3	74.7	45.5	96.9
1979	137.5	86.5	51.0	110.4
1980	174.4	118.3	56.1	119.5
1981	175.1	112.4	62.7	128.0[a]

[a]Updated using Eisner and Pieper's methodology.

value of government land from the Federal Reserve Board's estimate of the market value of private land, and from the ratio of Milgram's 1968 estimate to the 1968 Federal Reserve Board's estimate for private land. Thus,

$$GL_t = PL,F_t \cdot (GL,M_{68}/PL,F_{68}),$$

where GL_t is the market value of government land at t, PL,F_t is the Federal Reserve Board's estimate of the market value of private land in year t, and GL,M_{68} is Milgram's estimate of the market value of federal land in 1968. Table 6 reproduces the estimates of these authors for selected years.

Eisner and Pieper's estimates for the value of federal land are consistent with recent work by Goldsmith (1982) on the national balance sheet. Milgram's study provides estimates of the value of land over all sectors, 1952–68. Goldsmith extrapolates these estimates to 1975, arriving at a figure for the aggregate value of land of $1551 billion. He estimates the share of federal land at 4 percent, which would give an estimate of $62 billion for the value of federal land in 1975, compared with Eisner and Pieper's $63.6 billion.[21]

B. *New Estimates of the Value of Federal Land*

We have extended Milgram's estimates of the value of federal land to the period 1969–81, using a variant of her methodology as described in the Appendix. In Table 7, these estimates can be compared with Eisner's extension of the Goldsmith/Milgram estimates.

We estimate the value of federal land to be $175 billion in 1981, composed of $112 billion urban land and $63 billion rural land. Our new estimates substantially exceed those of Eisner and Pieper. The total is larger in every year, and the rate of growth is significantly higher.

Our estimates are higher because we take into account the change in the composition of federal land holdings. Eisner and Pieper

[21]Based on Goldsmith's estimates of total national assets in 1980 of $21,645 billion (1982, Table 89, p. 200), we can derive his estimate of the value of federal government land in that year. According to Goldsmith, land values constituted 13.7 percent of total national assets. Assuming that the federal government still holds 4 percent of total land values, the value of federal government land in 1980 would be $118.6 billion. This is again roughly the same magnitude as Eisner and Pieper's corresponding estimate of $124.9 billion and far smaller than our estimate of $174.4 billion. Goldsmith's estimates suggest that he, like Eisner and Pieper, does not take into account the substantial change in composition of federal government land.

TABLE 8—COMPOSITION OF FEDERAL LAND HOLDINGS SELECTED YEARS
(Millions of Acres)

Year	Urban	Rural	Total
1968	1406.3	753938.5	755344.8
1972	2326.8	758349.6	760676.4
1976	2936.2	759256.2	762192.4
1980	4768.4	714753.4	719521.6

do not consider the composition; their extension simply indexes the value of government land by the change in value of private land in the aggregate. While the total acreage held by the federal government declined by almost 5 percent between 1968 and 1981, its holdings of more valuable urban acreage more than tripled. This is shown in Table 8.

This total value estimate of $175 billion is approximately the value of all of the equipment (such as machines, trucks, typewriters, computers, etc.) owned by the federal government (as estimated by Eisner and Pieper). It is far less than the value of federal mineral rights as estimated in Section I, Part C.

It may be tempting simply to add the value of federal land to mineral rights. However, there are theoretical reasons to suspect that at least some of the value of federal mineral rights is capitalized into the value of the federal land bearing the minerals. In that case, it would be necessary to discount the value of the land before aggregating the land and the mineral rights. Since we are uncertain of the extent of capitalization, we have focused on the disaggregated components rather than their sum.

We have clearly not captured the full value of federal onshore mineral rights in our land value figures. Our estimate of the value of onshore mineral rights in 1981 ($146 billion) greatly exceeds the value of federal rural land ($63 billion). Furthermore, the method of derivation for estimates of federal rural land values makes it unlikely that they will reflect the underlying mineral values.[22] Clearly, to obtain the total value of land and mineral rights, we would have to add some, though not all, of the value of federal land to our estimate of the value of federal mineral rights.

III. Summary and Conclusion

We have presented new and updated estimates of the value of federal land and oil and gas mineral rights. These estimates are $175 billion and $819 billion, respectively, by 1981.[23]

Our results reveal the increasing importance of the value of the federal government's holdings of urban land in the total value of federal land over the period from 1968 to 1981. Over this interval the federal government's holdings of urban land tripled in acreage.

We estimate the value of federal oil and gas mineral rights to be very substantial, particularly following the sharp rise in energy prices in 1974 and 1979. In 1981, for example, the magnitude exceeds the privately held national debt.

This study is unique, as far as we know, in its effort to determine the value of the federal

[22] The magnitude of the rural land value estimates is largely determined by the average estimated price of the public domain managed by the BLM. (See the Appendix.) The BLM price estimates do not appear to take mineral rights into account.

[23] Recently, the Minerals Management Service (MMS) (1985) estimated the value of undiscovered offshore oil and gas to be $95 billion, one-sixth our estimate. The difference is primarily attributable to two sources: 1) lower estimates of quantities due to revision of estimates, leasing, and discoveries between 1981 and 1986; and 2) differences in value per barrel due partly to the recent fall in oil prices and partly to MMS assuming price growth of 1 percent per year, but real interest rates of 8 percent.

government's mineral rights and in the methodology used. We exploit information about undiscovered reserves and royalty and bonus payments to the government. Our methodology can be extended both to other minerals and to the private sector.

There remains room for considerable research on both the value of government land and its mineral rights. A new benchmark estimate for the value of federal land in a particular year is especially important. A more detailed disaggregation of the types of federal land holdings and improved and updated corresponding price information would also be helpful.

Improved estimates of the size of other mineral resources on federal land and estimates of the relationship between the quantity of economically recoverable unproven reserves of oil and gas and their prices would enable us to produce more comprehensive measures of wealth.

Finally, let us reemphasize the potential importance of estimates such as ours to sensible government budgetary decisions. These include land management policy, general cost-benefit analysis incorporating proper measures of the opportunity cost of resources, and perhaps even, in some contexts, fiscal policy.

APPENDIX

Milgram's time-series estimates, z_{56}, for the period 1952–68 rely on Goldsmith's estimate of the value of federal land (1962, p. 188). She uses the following price indexes:

A nonmetropolitan price index, IR_i, based upon the value of farmland per acre, as estimated by the U.S. Department of Agriculture.

An urban price index, IU_i, based upon three component indexes: FHA site prices of residential land values in Los Angeles and Philadelphia.

Average price per acre of rural land held by the federal government, PR_i. This is a weighted average of two indexes: the value of the public domain in the jurisdiction of the Bureau of Land Management (BLM) (90 percent), and the value of farmland per acre (10 percent).

She uses three additional statistics from the BLM:

An estimate of the value of government land, LC_i.

Acreage held by the federal government, broken down as rural $= AR_i$, and urban AU_i.

The statistics of the BLM appear in *Summary Report of Real Property Owned by the United States Throughout the World*, General Services Administration.

Milgram computes three series: value of rural land $= MR_i$; value of urban land $= MU_i$; total value of federal government land $= MA_i$.

An algebraic expression of Milgram's procedure is

$$MA_t \equiv MR_t + MU_t$$

$$MR_{52} = (PR_{56} \cdot AR_{56})$$

$$MR_t = MR_{t-1} \cdot \frac{IR_t}{IR_{t-1}} + PR_t \cdot \Delta AR_t$$

$$MU_t = MU_{t-1} \cdot \frac{IU_t}{IU_{t-1}} + \Delta LC_t - PR_t \cdot AR_t,$$

where $\Delta LC_t = LC_t - LC_{t-1}$.

Comments: 1) Milgram's estimates for 1952–55 and 1967–68 are extrapolations of the series determined by the method above. 2) PR_t could not be determined for our extension from Milgram's sources. Our extension from 1969 on uses the published farmland series, but a series on the value of the public domain was not available from the BLM. Therefore, the public domain component of the weighted average is based on an extrapolation of the value of the public domain in 1968. The growth rate for each subsequent year was assumed to equal the average annual growth rate of farmland over the period 1969–81.

REFERENCES

Boskin, Michael J., "Federal Government Deficits: Myths and Realities," *American Economic Review Proceedings*, May 1982, 72, 296–303.

Eisner, Robert, "Capital Gains and Income: Real Changes in the Value of Capital in the United States, 1946–1977," in Dan Usher ed., *The Measurement of Capital*, Chicago: University of Chicago Press, 1980, 175–346.

______ **and Pieper, Paul J.,** "A New View of the Federal Debt and Budget Deficits," *American Economic Review*, March 1984, *74*, 11–29.

Goldsmith, Raymond W., *The National Wealth of the United States in the Postwar Period*, Princeton: Princeton University Press, 1962.

______, *The National Balance Sheet of the United States, 1953–1980*, Chicago: University of Chicago Press, 1982.

Landefeld, J. Steven and Hines, James R., "Valuing Nonrenewable Natural Resources: The Mining Industries," in *Measuring Nonmarket Economic Activity*, Working Paper 2, Bureau of Economic Analysis, U.S. Department of Commerce, 1982.

Milgram, Grace, "Estimates of the Value of Land in the United States Held by Various Sectors of the Economy, Annually, 1952 to 1968," (Appendix II) in Raymond Goldsmith, ed., *Institutional Investors and Corporate Stock*, New York: Columbia University Press, 1973.

Reeve, J. E. et al., "Government Component in the National Wealth," in *Studies in Income and Wealth*, No. 12, New York: National Bureau of Economic Research, 1950.

Soladay, John J., "Measurement of Income and Product in the Oil and Gas Mining Industries," in Dan Usher, ed., *The Measurement of Capital*, Chicago: University of Chicago Press, 1980.

American Petroleum Institute, *Basic Petroleum Data Book, Petroleum Industry Statistics*, Vol. 4, No. 1, Washington, 1984.

Financial Accounting Standards Board, *Statement of Financial Accounting Standards No. 39: Financial Reporting and Changing Prices: Specialized Assets Mining and Oil and Gas*, Stamford: Financial Accounting Standards Board, October 1980.

United Nations Economic and Social Council, Statistical Office, *Future Directions for Work on the System of National Accounts*, New York, 1979.

______, *Future Directions for Work on the System of National Accounts*, New York, 1980.

U.S. Council of Economic Advisors, *Economic Report of the President*, Washington: USGPO, 1984.

U.S. Department of the Interior, Minerals Management Service, *Federal Offshore Statistics*, December 1983.

______, *Mineral Revenues*, December 1983.

______, "Five-Year Leasing Program for the Outer-Continental Shelf, 1986–90," Appendix F, 1985.

U.S. Geological Survey, *Estimates of Undiscovered Recoverable Conventional Resources of Oil and Gas in the United States*; Circular 860, Washington, 1981.

U.S. Securities and Exchange Commission, "Oil and Gas Producers' Supplemental Disclosures," *Federal Register: Rules Regulations*, 1979.

The Simple Macroeconomics of Profit Sharing

By MARTIN L. WEITZMAN*

This paper is in the spirit of the "temporary equilibrium" approach to macroeconomics. It basically extends that framework to cover a profit-sharing system and then compares the macroeconomic characteristics with those of the more familiar wage system. A first, preliminary step is to demonstrate how a microeconomic model of monopolistic competition can be built up into a junior member of the Keynesian macro family. The methodology is to create from first principles —including a careful formulation of a monopolistically competitive product market structure—a natural underpinning for the standard aggregate demand specification.[1]

The primary goal of the paper is to apply the integrated monopolistic-competition-Keynesian-type apparatus described above to investigate the macroeconomic properties of a profit-sharing economy. The existence of a consistent general framework covering both cases invites meaningful comparisons that indicate clearly why an economy based on profit-sharing principles possesses natural immunity to stagflation. By contrast, the wage economy—a system we have largely accepted without critically examining its macroeconomic consequences—is more prone to suffer from unemployment and inflation. The policy implications for aggregate demand management in wage and share systems are analyzed and contrasted.

In writing this paper, my philosophy has been to not shirk from using those reasonable parameterizations and functional forms which yield nice crisp results and permit me to focus sharply on the essential logic of basic issues. It is certainly possible to present the main results in a somewhat more general formulation (as the astute reader will appreciate), but, I fear, only at some cost of distracting attention from those central features I wish to highlight.

I. The Demand Side

The stylized economy under consideration consists of three types of representative agents. The first type of agent is a producer or firm. There are n firms, each of which produces a different good, indexed $i = 1,2,\dots,n$, where n is taken to be a given large number.[2] A second class of agents is the households, of which there are a gigantic number, indexed $h = 1,2,\dots,H$, where $H \gg n \gg 0$. An autonomous government sector, the third agent, makes purchases, taxes households, and has an exclusive franchise on the creation of money.

*Department of Economics, MIT, Cambridge, MA 02139. I am indebted to Robert Solow for his detailed, useful comments on an earlier draft, to James Meade and Hal Varian for their helpful suggestions, and I thank an anonymous referee for critical comments. They should not be saddled with opinions or errors of the paper. The research was supported by a grant from the National Science Foundation.

[1] This may be a useful exercise by itself because, in my opinion, any macroeconomic framework is misleading without an underlying model of the firm based upon imperfect competition. For an elaboration of this view, see my 1982 and 1985 papers, Robert Solow (1985), or James Meade (1984). While some contributors to the important market disequilibrium school have attempted to cope with imperfect competition in the product market, I think it is fair to say that the issue has not been addressed directly and developed from first principles in the spirit of the present paper—using the "actual," rather than "perceived" or "conjectural," demand curves. For an admirable survey of the temporary fixed price approach, see Jean-Pascal Benassy (1982) and the references cited there. Aside from the emphasis on dealing with monopolistic competition from first principles, the rest of this paper's framework is similar to what is adopted in much of the fixed-price literature, although that approach, so far as I know, has never been used to analyze profit sharing.

[2] Behind the fixed number of firms are suppressed or suspended some interesting and important issues regarding barriers to entry or exit, economies of scale, sunk costs, irreversible investments, and the like. Some hint of what might be appropriate to a longer-run analysis is contained in the already cited articles by myself (1982) and by Solow.

There are three categories of commodities in the prototype economy. The first category consists of the n goods produced by the n firms. Goods are considered to be highly perishable, so that inventories are negligible and sales are always very nearly equal to production. Labor, the second category, is a homogeneous commodity inelastically supplied by the households. Money, the third kind of commodity, is storable, not producible by private agents, and can be costlessly created by the government. Money serves as the exclusive unit of account, medium of exchange, and store of value in the economy.

The production of good i, denoted Y_i, and its price, P_i, are of course chosen by firm i. The eventual analysis of that choice will constitute an ultimate aim of the paper. But, for the time being, suppose that prices are viewed parametrically by buyers, who act as if they can purchase as much as they want of any good at the prevailing prices $\{P_i\}$. As might be expected under monopolistic competition, it turns out that prices will always be chosen by firms so customers can buy as much as they want, and in that sense the product market always clears.[3]

Households obtain utility from consuming goods and holding money balances. The utility of money is indirect; it serves as a proxy for the value of future consumption goods that can be purchased when money is carried over into later periods. For simplicity, each household is postulated to have the *same* utility function. When a household consumes goods $\{C_i\}$ and holds money balances M, it obtains utility according to the expression:[4]

$$(1)\quad U(\{C_i\}, M/P) = \left(\left[\sum C_i^{(E-1)/E}\right]^{E/(E-1)}\right)^{\theta}\left(\frac{M}{P}\right)^{1-\theta}.$$

The aggregate price level P in the above expression is defined by the formula:

$$(2)\qquad P \equiv \left(\sum P_i^{1-E}/n\right)^{1/(1-E)},$$

which is the appropriate goods price index, from duality theory, to use for the postulated utility function (1).[5]

Formula (1) is a compound Cobb-Douglas utility function (with parameter θ, $0<\theta<1$), whose two arguments are money and a *CES* composite subutility function of goods. The elasticity of substitution between money and the composite good is unity, whereas the elasticity of substitution among the n goods is $E>1$.

With a current budget of B^h, household h confronts the problem

$$(3)\qquad \max U(\{C_i\}, \tilde{M}/P)$$

subject to

$$(4)\qquad \sum P_i C_i + 1\cdot \tilde{M} = B^h.$$

For a modified Cobb-Douglas utility function of the form (1), the solution to the above problem is

$$(5)\qquad \tilde{M}^h = (1-\theta)B^h,$$

$$(6)\qquad C_i^h = (P_i/P)^{-E}\theta B^h/nP.$$

The total amount of good i consumed in the economy is

$$(7)\qquad C_i \equiv \sum_h C_i^h$$

and aggregate consumption C may be consistently defined as

$$(8)\qquad C \equiv \sum P_i C_i/P.$$

[3] Indeed, I consider it a deep-seated characteristic of capitalism that the product market is practically always in a state of excess supply. See my 1984 book, ch. 3.

[4] Money in the utility function (1) serves as a link between the present and an uncertain future, with θ parameterizing the desire to consume now. There is an implicit presumption that the future can be collapsed into a dynamic-programming state-evaluation function like (1). On this point, see Benassy (pp. 87–88) or Jean-Michel Grandmont (1983, pp. 17–32).

[5] See, for example, Hal Varian (1984) or Avinash Dixit and Joseph Stiglitz (1977).

The following relations then hold:

$$C = \theta B/P \tag{9}$$

$$\tilde{M} = (1-\theta)B \tag{10}$$

$$\tilde{M} + PC = B \tag{11}$$

$$C_i = (P_i/P)^{-E} C/n \tag{12}$$

where

$$B \equiv \sum B^h \tag{13}$$

$$\tilde{M} \equiv \sum \tilde{M}^h. \tag{14}$$

Total government real spending on goods, denoted A, is treated as autonomously determined. The government's tradeoff among goods is considered, for convenience, to be the same as the household's, given by the utility function:

$$V(\{A_i\}) = \left(\sum A_i^{(E-1)/E}\right)^{E/(E-1)}. \tag{15}$$

The government maximizes (15) subject to the budget constraint

$$\sum P_i A_i = PA \tag{16}$$

which yields the solution

$$A_i = (P_i/P)^{-E} A/n. \tag{17}$$

Aggregate demand for good i by the consumers and the government is

$$Y_i \equiv C_i + A_i. \tag{18}$$

With aggregate real output defined as

$$Y \equiv \sum P_i Y_i / P, \tag{19}$$

definition (18) yields

$$Y = C + A \tag{20}$$

(from combining with (8) and (16)), and

$$Y_i = (P_i/P)^{-E} Y/n \tag{21}$$

(from combining with (12) and (17)).

The government collects the fraction s of each household's current income as taxes. National income is PY, all of which is distributed to households as wages plus profits. Aggregate disposable income is therefore

$$PY_d = (1-s)PY \tag{22}$$

and the total budget of all households is

$$B = (1-s)PY + M \tag{23}$$

where M represents the aggregate stock of money initially held by all households at the beginning of the period under consideration.

It follows directly from (23), (11), and (20) that

$$PA - sPY = \tilde{M} - M, \tag{24}$$

that is, the government finances its deficits by inducing households to hold more money.

Using (9) and (23) to eliminate B gives

$$C = \theta(1-s)Y + \theta(M/P) \tag{25}$$

which is the relevant aggregate consumption function for the economy, with $\theta(1-s)$ the marginal propensity to spend out of income.

Combining (20) with (25) yields

$$Y = \alpha A + \beta(M/P) \tag{26}$$

where

$$\alpha \equiv 1/(1-\theta(1-s)) \tag{27}$$

$$\beta \equiv \theta/(1-\theta(1-s)) \tag{28}$$

are the relevant fiscal and monetary multipliers.

Equation (26) can be interpreted as a reduced-form Keynesian-type macroeconomic relation. Strictly speaking, monetary policy (as that term is usually understood) does not have an independent role to play in the current formulation because no distinction is being made between monetary and other financial assets or operations. But I feel that a simplistic association of M with the "stock of money" (and of open market operations with "money rain"), conveys the spirit of

what a more sophisticated analysis might prove rigorously. Although I have found it valuable to think in terms of an integrated micro-macro framework developed from what I view as first principles, it is possible to treat (26) simply as a behavioral relationship having the traditional *IS-LM* interpretation.

Condition (26) is the fundamental macroeconomic equation of the paper, summarizing all relevant information about aggregate demand given only that buyers are able to purchase whatever goods they want at prevailing prices.

II. Prices and Production

It is important to realize that the Keynesian demand specification (26) typically forms an underdetermined system. Given A and M (and the parameters α and β), equation (26) describes a relation that must hold between two macroeconomic variables: Y and P. The traditional procedure for making the system determinate is to postulate a fixed price level

$$P = \overline{P} \tag{29}$$

for the short run.[6] In this paper I want to *derive* (29) as the profit-maximizing response of a large number of monopolistically competitive firms constrained to pay fixed money wages. The same methodology will then be applied to the case where the fixed contract is of a profit-sharing form, which will yield quite different solution properties and macroeconomic implications from (29).

Suppose that each of the n different goods is produced by the same production technology. Firm i $(1 \leq i \leq n)$ produces Y_i units of good i from L_i employees according to the formula

$$Y_i(L_i) = \gamma(L_i - f), \tag{30}$$

where γ is the marginal productivity of an extra worker and f represents a fixed amount of overhead labor which must be employed to produce any output at all. The production function (30) can be viewed as a first-order approximation in the relevant operating range.[7]

The total amount of labor employed is then

$$L \equiv \sum L_i. \tag{31}$$

If L^* represents the total available labor, assumed to be inelastically supplied by households, then the condition

$$L \leq L^* \tag{32}$$

must be obeyed in the aggregate.[8]

In any symmetric situation, aggregate output must be given by the formula

$$Y = \gamma(L - F), \tag{33}$$

where

$$F \equiv nf. \tag{34}$$

From (30), then,

$$Y \leq Y^* \tag{35}$$

where

$$Y^* \equiv \gamma(L^* - F) \tag{36}$$

represents potential aggregate output.

What follows in this section is an overview of the methodology to be followed in analyz-

[6]An alternative is to postulate an "aggregate supply function" which is, I feel, a dubious macroeconomic concept at best, especially for a world where firms are price makers in imperfectly competitive product markets.

[7]That unit variable costs are roughly constant over some range is, I think, a decent enough stylized fact to be used as a point of departure for the purposes of this paper.

[8]The reader who wants to should be able to redo the analysis of this paper for the case where labor supply is not perfectly inelastic. Nothing of substance changes. In long-run equilibrium, wage and profit-sharing systems will continue to be identical. In the short run, when pay parameters are sticky, a profit-sharing economy effectively banishes involuntary unemployment, while a wage economy may have it, even in the presence of elastically supplied labor. The message is essentially the same as when labor is perfectly inelastic.

ing the short-run price and production decisions of the firms. Suppose the cost per worker of hiring L_i workers is $W(L_i)$, where the average pay function $W(\cdot)$ is exogenously given in the short run and is identical for each firm. The relevant equilibrium concept is taken to be a symmetric Nash equilibrium in prices. Each firm charges an identical price, which is the profit-maximizing price for it given that all other firms are charging that same price. The corresponding output and employment decisions are those needed to support the profit-maximizing Nash equilibrium behavior.

A short-run macroeconomic equilibrium is a price P, aggregate output level Y, and total employment L simultaneously satisfying (26), (32), (35), and the conditions

$$P(Y/n)-W(L/n)\cdot L/n = \max_{P_i, Y_i, L_i}\{P_iY_i - W(L_i)\cdot L_i\}, \tag{37}$$

subject to

$$L_i \le L^* - (n-1)L/n \tag{38}$$

$$Y_i \le \gamma(L_i - f) \tag{39}$$

$$Y_i \le (P_i/P)^{-E}Y/n. \tag{40}$$

It is easy to verify that any solution of the constrained optimization problem (37)–(40) will satisfy (39) and (40) with strict equality. Since (40) is ultimately derived from consumer demand conditions, when it holds with full equality, buyers are able to purchase whatever they want at prevailing prices and, hence, in the aggregate (26) must be satisfied.

So long as n is a large number, each firm i is justified in regarding its demand Y_i, given by (21), as a true function of only its own price P_i, with aggregate variables P and Y parametrically fixed beyond its control.[9]

[9]This statement can be rigorously defended.

III. Short-Run Equilibrium in a Wage Economy with a Parametrically Given Wage

In the short run, suppose each firm i pays labor an exogenously fixed money wage

$$W(L_i) = w \tag{41}$$

where w is treated as autonomously given.

The state of the macroeconomy is described by the basic aggregate demand equation (26). The extra degree of freedom in (26) between the variables Y and P is determined by firms' profit-maximizing Nash equilibrium behavior (37)–(40) given the rigid wage (41).

Let

$$\mu \equiv E/(E-1) \tag{42}$$

be the markup coefficient for each firm. The coefficient μ represents the ratio of average revenue (price) to marginal revenue.

With the production function (30) and the labor payment schedule (41), the marginal cost of an extra unit of output to firm i is w/γ. For the demand function (21), marginal revenue at a price of P_i is P_i/μ. Hence, if availability of labor were not a binding constraint, each firm i would choose to set a price

$$P_i = \mu w/\gamma \tag{43}$$

and the desired or target output of the wage system, denoted $\hat{Y}$, would then be, from the aggregate demand condition (26):

$$\hat{Y} \equiv \alpha A + \beta M\gamma/\mu w. \tag{44}$$

Define the tautness or tension of the wage system as

$$\tau \equiv \hat{Y} - Y^*. \tag{45}$$

The variable τ measures the difference between desired output (what firms would like to produce in the aggregate on the given wage contract if there were no overall labor constraint) and potential output (what the system is physically capable of producing). $[\tau > 0]$ is a region of positive excess demand

TABLE 1—SHORT-RUN BEHAVIOR OF MAJOR MACROECONOMIC VARIABLES IN A WAGE SYSTEM

Variable	$\tau<0$	$\tau>0$
Y	$\alpha A+\dfrac{\beta M\gamma}{\mu w}$	Y^*
P	$\mu w/\gamma$	$\dfrac{\beta M}{Y^*-\alpha A}$
W/P	γ/μ	$\dfrac{w(Y^*-\alpha A)}{\beta M}$

for labor, whereas $[\tau<0]$ is a region of negative excess demand for labor.

The unique symmetric Nash equilibrium with each firm playing its own price as a profit-maximizing strategy given the fixed wage (41) depends on the underlying configuration of parameters. Equilibrium values of the major macroeconomic variables are shown in Table 1.

That Table 1 describes the unique symmetric Nash equilibrium of a fixed-wage economy should be fairly clear. Condition (43) has already been explained for the case where the firm can buy as much labor as it wants at the fixed wage (41). The corresponding value of Y in region $[\tau<0]$ follows immediately from (26).

In the excess demand for labor region $[\tau>0]$, aggregate output must be at its maximum feasible amount Y^*, with the corresponding value of P determined from (26). That such a configuration represents a Nash equilibrium in prices is easily verified. Since the marginal revenue product of labor exceeds the marginal cost of labor when $\tau>0$, the firm would like to reduce its price and to produce more output, if only it could find more labor to hire. With each firm's output level effectively constrained (by (38), (39)) to be no more than Y^*/n in the case $\tau>0$, it is unprofitable for a firm to lower price unilaterally, and it certainly is not profitable to restrain output further by raising price. (Of course, firms could also increase the money wage to attract more workers, and will do so in the long run, but this has been ruled out as short-run behavior by assumption.)

From Table 1, the macroeconomic properties of a fixed-wage economy depend essentially on whether the system is in a state of positive or negative tension. It is important to understand fully the meaning and significance of this dichotomy, because the same logic will carry over—albeit with an important and unexpected twist—to analyzing the short-run behavior of a profit-sharing economy.

The profit-maximizing response to demand changes of a monopolistically competitive firm facing an isoelastic demand curve and constant marginal cost is to charge the same price and vary production accordingly. A Nash equilibrium of such firms with fixed money wages satisfying the condition $\tau<0$ yields the familiar fixed-price world of Keynesian "underemployment equilibrium."

In such a world, prices are basically set by producers as a direct markup over wages independent of the state of aggregate demand. From formula (43), the coefficient of proportionality between P and w is μ/γ. So it is a fair approximation to treat prices as proportional to unit labor costs in underemployment states of a fixed-wage economy—provided there is no systematic tendency for the markup coefficient divided by the marginal productivity of labor to vary significantly over the business cycle.[10]

A fixed-wage economy in region $[\tau<0]$ exhibits textbook Keynesian behavior in the short run. P cannot be directly affected by government policy, but Y and L respond via the standard Keynesian multipliers to changes in A, M, or s.

By contrast, a fixed-wage economy in the region $[\tau>0]$ displays classical or monetarist characteristics. Government aggregate demand management has no influence on real output, already at full employment, *but directly and powerfully influences the price level.* Monetary policy is strictly neutral, with

[10] Note that the main conclusions come from the near constancy of the ratio μ/γ, not from the separate constancies of μ and γ. The model and its basic implications would not be significantly altered if elasticities and marginal costs were allowed to vary systematically in such a way that the ratio μ/γ remained unchanged. Sidney Weintraub (1981) drew attention to the important empirical regularity of a near-constant average markup of prices over unit labor costs. See Weintraub and references to other works there cited.

prices proportional to M. Expansionary fiscal policy has only an inflationary impact, since it crowds out private spending.

Summing up, then, there is a kind of abstract symmetry in the short-run behavior of a fixed-wage economy. With $\tau < 0$, government policy is effective at altering real economic activity, but ineffective at changing prices. When $\tau > 0$, government policy is effective in determining the price level, but ineffective at influencing real aggregate variables. While the demarcation between the two regimes is unlikely to be nearly as clear cut in practice as in theory (partly because wages are more flexible upward than downward), I nevertheless feel the distinction is conceptually useful.

IV. Long-Run Equilibrium in a Wage Economy with a Competitively Determined Wage

Consider a longer-run situation where everything is as described in the previous section, only now the wage is endogenously determined by thoroughgoing competition in the labor market. Under competition, each firm is free to set its own wage rate, and will do so to maximize profits taking as given the prevailing level of pay throughout the economy. The limiting Nash equilibrium behavior (as each firm becomes a negligible buyer of labor) yields the full-employment wage at which the marginal revenue product of labor is everywhere equal to the uniform rate of pay and the sum of labor demands just equals the supply of labor. Each firm is then offering an identical wage, which is the profit-maximizing wage for it to offer given that all other firms are offering that same wage.

I should point out that I view the hypothesis of a competitive equilibrium wage not as a literal description of the state of the labor market, but more as an approximation or norm which is never actually attained yet forms a useful basis for talking about possible departures from normalcy. The "competitive wage" represents a long-term tendency which, on the one hand, cannot be indefinitely thwarted with impunity but, on the other hand, is unlikely to hold fully at any particular time or place because "other" variables are changing too rapidly and unpredictably.

The long-run competitive equilibrium wage, taking all else about the wage system as given by last section's description, is

$$w^* = \beta M\gamma/\mu(Y^* - \alpha A). \tag{46}$$

When $w = w^*$, there is no unemployment, and the demand for labor just equals the supply. Under competitive forces in the labor market, then, the wage system gravitates toward the region $[\tau = 0]$ of zero tautness which just divides the "Keynesian" $[\tau < 0]$ and "classical" $[\tau > 0]$ regions.

It follows that an economy whose long-term wage tendencies are described by (46) will display all of the neutrality and policy-ineffectiveness results of classical macroeconomics—in the long run. For example, changes in M will "eventually" generate equiproportionate changes in w, and hence in P, so that nothing real is altered in the economy.

While some long-run competitive forces are pushing a wage economy toward $[\tau = 0]$, they are unlikely to be decisive at any given time since the whole system is precariously balanced on the output side. The boundary region $[\tau = 0]$ is a very thin set, a razor's edge of measure zero, so it is extremely improbable that a capitalist wage economy should remain there for long. In fact the *realpolitik* of wage capitalism, with its less-than-perfect labor markets and downward-inflexible wages, has the system residing in region $[\tau < 0]$ most of the time, hopefully not too far from the full-employment boundary $[\tau = 0]$. It seems a fair empirical generalization to say that the relevant region for most short-term policy analysis is the Keynesian region $[\tau < 0]$ where

$$w > w^*. \tag{47}$$

V. Short-Run Equilibrium in a Profit-Sharing Economy with Given Pay Parameters

In the short run, suppose each firm i pays its workers by the profit-sharing formula

$$W(L_i) = \omega + \lambda\left(\frac{R_i(L_i) - \omega L_i}{L_i}\right), \tag{48}$$

where $R_i(L_i)$ stands for total revenue as a function of labor, given the demand function (21) and the production function (30). The pay parameters ω, representing the base wage, and $\lambda > 0$, representing the profit-sharing coefficient, are both treated in the short run as exogenously fixed.[11]

The methodology for determining a short-run equilibrium in a profit-sharing economy is exactly the same as in a wage economy. The profit-sharing firm makes its short-run pricing, output, and employment decisions to maximize profits given the rigid labor payment formula (48) and given the prices that all of the other firms are charging. The economy's short-run behavior is modeled as the Nash equilibrium outcome, (37)–(40), of this individualistic profit-maximizing process which simultaneously satisfies the basic macroeconomic condition (26).

The wage bill if L_i workers are hired by firm i is, from (48),

$$(49)\quad W(L_i)\cdot L_i = (1-\lambda)\omega L_i + \lambda R_i(L_i)$$

and net profits are

$$(50)\quad \pi_i(L_i) \equiv R_i(L_i) - W(L_i)\cdot L_i.$$

Combining (49) with (50), the net profits of firm i can be rewritten in the form

$$(51)\quad \pi_i(L_i) = (1-\lambda)(R_i(L_i) - \omega L_i).$$

If unlimited amounts of labor are available to be hired on the share contract (48), from (51), the firm will choose to hire workers to the point where

$$(52)\quad R_i'(L_i) = \omega.$$

But the marginal revenue product of labor with demand curve (21) and production function (30) is related to price charged, P_i, by the formula

$$(53)\quad R_i'(L_i) = \gamma P_i/\mu.$$

Combining (52) and (53), with unlimited supplies of labor available on the pay schedule (48), each firm i would choose to set its price at the level

$$(54)\quad P_i = \mu\omega/\gamma.$$

The corresponding desired or target aggregate output level of the profit-sharing system with fixed-pay parameters (ω, λ), denoted $\hat{Y}'$, would then be, from (26),

$$(55)\quad \hat{Y}' = \alpha A + \beta M\gamma/\mu\omega.$$

The hypothetical variable $\hat{Y}'$ measures what firms would like to produce in the aggregate on the given pay contract if there were no overall labor constraint.

The tautness of the profit-sharing system is then

$$(56)\quad \begin{aligned}\tau' &\equiv \hat{Y}' - Y^* \\ &\equiv \alpha A + \beta M\gamma/\mu\omega - Y^*\end{aligned}$$

Note that the degree of tautness varies inversely with ω, and that a "pure" sharing system not having any base wage would possess an infinite demand for labor.

The unique symmetric Nash equilibrium with each firm setting its own price at a profit-maximizing value given all other firms' prices, and given the fixed profit-sharing pay formula (49), depends on the underlying configuration of parameters as shown in Table 2.

The reasoning to explain why Table 2 describes the unique symmetric Nash equilibrium of a profit-sharing economy closely parallels the reasoning behind Table 1 and is omitted here for the sake of brevity. In both cases the key insight is that actual aggregate output must be the smaller of a demand-determined target and a supply-determined capacity. The rest follows directly.

The most immediately striking thing about Table 2 is that the first two rows are exactly

[11] The above formulation omits intermediate materials, mostly for the sake of simplicity. While there may be some practical problems with profit sharing due to the fact that, in the real world, "profits" is a somewhat elastic concept, I do not see insurmountable difficulties arising here. In any event, treatment of such considerations (and also bankruptcy, legal issues, leverage effects, etc.) is well beyond the scope of the present paper.

TABLE 2—SHORT-RUN BEHAVIOR OF MAJOR MACROECONOMIC VARIABLES IN A PROFIT-SHARING SYSTEM

Variable	$\tau' < 0$	$\tau' > 0$
Y	$\alpha A + \dfrac{\beta M \gamma}{\mu \omega}$	Y^*
P	$\mu\omega/\gamma$	$\beta M/(Y^* - \alpha A)$
W/P	$(1-\lambda)\dfrac{\gamma}{\mu} + \lambda \dfrac{Y}{L}$	$(1-\lambda)\dfrac{\omega(Y^* - \alpha A)}{\beta M} + \lambda \dfrac{Y^*}{L^*}$

the same as in Table 1 except for ω replacing w. The share parameter λ does not affect real national product or the price level.

When firms are maximizing a function of the form (51), their reactions are not influenced by λ. So long as spending behavior is postulated to depend only on the level of aggregate income, and not its distribution, the pricing and output decisions of firms in any short-run equilibrium of the system must be independent of λ. The particular case $\lambda = 0$ is just the wage economy, which accounts for the near-identity between the first two rows of Tables 1 and 2. While values of λ affect the distribution of national income, they do not influence its determination. Only the value of ω, representing to a firm the "hard" money cost of taking on an extra worker (as opposed to the "soft" cost of a share of incremental gross profits), influences the overall level of national income. If workers in a wage economy agree to receive 80 percent of their pay in the form of base wages and 20 percent in the form of a profit-sharing bonus, the effect on national product, employment, and prices is "as if" wages had been cut by 20 percent while aggregate demand was being maintained at the same level.

When a wage economy suffering from unemployment converts to a profit-sharing formula whose parameters are initially set so that each employed worker is at first paid the same amount, the change will make all workers better off after adjustment. From (48), the real pay in a profit-sharing system is

$$W/P = (1-\lambda)\omega/P + \lambda(Y/L). \tag{57}$$

After conversion from a wage system to an "equivalent" profit-sharing system initially yielding the same pay, the share economy expands output and employment while lowering price. (Compare Tables 1 and 2 when $\omega < w$.) If labor productivity does not behave countercyclically ($F \geq 0$), from (57) real pay must increase.[12] In addition, new jobs have been created, so there are more employed workers, each of whom is receiving higher real pay. In this sense a move towards profit sharing represents an unambiguous improvement for the working class.

Note that the argument applies only when all (or almost all) firms of a wage economy simultaneously convert to profit-sharing plans. If one firm alone converts, it will hire new workers, but at the expense of driving down the pay of its original workers. So coordination may be required to induce people to convert to a share system; one possibility is to have the government reward profit-sharing workers, by preferential tax treatment of share income, for their part in creating the positive externality of a tight labor market.[13]

[12]Actually, all that is needed is that $\lambda Y/L$ not decrease faster than $(1-\lambda)\omega/P$ increases.

[13]I do not currently have a precise formulation of the "positive externality of a tight labor market" that could serve as an operational framework for analysis. Nevertheless, it seems intuitively clear to me that there may be a basic problem of institutional instability in a profit-sharing economy because high-λ behavior that is socially rational may not be individually rational. Some preliminary thoughts on this point are expressed in my 1984 book, ch. 9. I believe the relevant externality has to do with the idea that high stable pay for "insider" workers of the existing labor force (at the expense of "outsider" unemployed workers and the young) suits the interests both of the high-seniority employed workers

Comparing the first two rows of Table 2 with Table 1, the short-run aggregative properties of wage and share systems appear to be very analogous, the only essential difference being in the values of the variables w and ω. That interpretation is true, but it is deceptive, as will be shown presently.

VI. Long-Run Equilibrium in a Profit-Sharing Economy with Competitively Determined Pay Parameters

Consider next a longer-run situation where the setup is the same as in the last section, except that pay parameters are endogenously determined by thoroughgoing competition in the labor market. The basic concept of competitive equilibrium in the labor market is essentially the same for a share system as for a wage system. Given the pay parameters every other firm is selecting, each firm is free to choose its own pay parameters but must live with the consequences of labor shortage if it selects too-low values. The underlying solution concept is a symmetric Nash equilibrium in pay parameters, which means that if all firms are selecting (ω, λ) as parameter values, it is not profitable for any one firm to deviate from that pattern. This equilibrium value will be used primarily as a reference point to indicate the approximate region in pay-parameter space where a profit-sharing system is likely over time to end up.

A basic theoretical result to be proved below is that any pair $(\omega, \lambda) > 0$ constitutes a long-run competitive equilibrium in pay parameters if and only if it delivers to each worker the same pay as an equilibrium wage system $(w^*, 0)$ operating under otherwise identical circumstances. From (48), such an equivalence can be written as

$$(58) \qquad w^* = \omega + \lambda((P^*Y^* - \omega L^*)/L^*),$$

where w^* is defined by (46) and

$$(59) \qquad P^* = \beta M/(Y^* - \alpha A).$$

There is thus an inverse relationship between long-run equilibrium values of λ and ω and, hence, one extra degree of freedom in determining the pay parameters of a profit-sharing system.

I do not have a formal theory that would explain: (a) why a society chooses a particular (ω, λ) configuration; or (b) why pay parameters are sticky in the short run. I only have consistent stories about viable long-term combinations of ω with λ, and about the short-term consequences of pay parameters being temporarily frozen at various values. This partly intuitionist, partly formalist approach strikes me as the best feasible way of addressing the important issues involved. (And, presumably, the present analysis would be needed anyway as a preliminary step toward any more ambitious formulation directly attempting to tackle (a) and (b) above.) In my story, it is perhaps conceptually useful to think of λ as a policy variable chosen by the government to automatically "stabilize" the macroeconomy at full employment.[14]

and of their satisficing employers (who are doing well enough to want to continue enjoying the benefits of a quiet life). Converting outsider nontenured workers into permanent insiders may require institutional changes in the incentive structure going far beyond anything in current official thinking. Strong material incentives, such as favorable tax treatment of the profit-sharing component of a worker's pay will probably be needed to convince senior workers to acquiesce in a profit-sharing scheme with no restrictions on new hiring. (For a more extensive discussion of the problem of new hires, see my 1984 book, pp. 108–109 and 132–34.) A formal development of such ideas is properly the subject of future research, the current paper being limited to describing the macroeconomic implications of wage and profit-sharing systems without yet attempting the grand historical synthesis of explaining how or why they actually come into being.

[14] The fact that we generally observe $(\omega, \lambda) = (w^*, 0)$ (i.e., no profit sharing) might be because $(w^*, 0)$ represents some sort of institutional Nash equilibrium, with other combinations of (ω, λ) not sustainable in the face of possible externality/free-rider problems. (On this, see the suggestive discussion of my 1984 book, ch. 9.) Although intuitively plausible, this interpretation remains speculative. If true, it might justify public policy to induce high values of λ. Note, however, that most private companies in the immensely successful economies of Japan, Korea, and Taiwan pay a very significant fraction of worker remuneration as a bonus which is, or so it seems in many instances, at least indirectly linked

Then, over a longer term, ω can be envisioned as adjusting to satisfy (58). Throughout the short run, in my scenario, ω and λ are both thought of as being quasi-fixed parameters.

The explanation of (58) is roughly as follows. In long-run competitive equilibrium, due to migration pressure, each worker must end up with the same pay no matter what is the ostensible form of the payment (how it is split between straight money wages and shares of profit). Given the fact that every firm must end up paying the prevailing pay whatever parameter values it selects, the profit-sharing firm can do no better in the long run than to hire labor to the point where the marginal revenue product of an extra worker is equal to the prevailing pay, then setting its pay parameters accommodatingly during contract time to yield that going compensation for its workers.

Solely to preserve neatness and to save on space, (58) will be proved here only for the case $\omega = 0$ (pure revenue sharing). The proof for the more general case is essentially identical, although made considerably messier due to the additional notation which is required.[15]

Let $L(\lambda; \lambda^*)$ stand for the amount of labor any firm is able to attract if it pays a share λ when all other firms are paying equilibrium shares λ^*. If every other firm is paying a share λ^*, and there are a large number of firms, the prevailing level of pay must be $\lambda^* P^* Y^*/L^*$ where, because any long-run equilibrium is at full employment, P^* is given by (59). It follows that $L(\lambda; \lambda^*)$ must satisfy the condition

$$(60) \qquad \frac{\lambda R(L(\lambda;\lambda^*))}{L(\lambda;\lambda^*)} = \frac{\lambda^* P^* Y^*}{L^*}$$

where $R(L)$ stands for a firm's revenue as a function of the labor working for it. Since (60) must hold for all λ, differentiating with respect to λ and collecting terms yields

$$(61) \qquad \frac{\partial L}{\partial \lambda} = R \Big/ \left(\frac{\lambda^* P^* Y^*}{L^*} - \lambda R' \right).$$

The long-run equilibrium problem of the firm, given λ^*, is to select λ to maximize $(1-\lambda)R(L(\lambda;\lambda^*))$, which yields the first-order condition

$$(62) \qquad (1-\lambda)R'(\partial L/\partial \lambda) = R.$$

Combining (61) with (62),

$$(63) \qquad R' = \lambda^* P^* Y^*/L^*.$$

But from (53), the marginal revenue product of labor for a firm equals γ/μ times its optimally chosen price. Hence there will be system-wide equilibrium if and only if

$$(64) \qquad (\gamma/\mu)P^* = \lambda^* P^* Y^*/L^*$$

or iff (from (46) and (59))

$$(65) \qquad w^* = \lambda^* P^* Y^*/L^*$$

which is exactly the condition (58) to be proved for the case $\omega = 0$.

There are two major implications of what has been derived in this section. The first is that wage and profit-sharing systems are isomorphic in a long-run stationary equilibrium with competitive labor markets. I take this to mean that both systems have some long-run tendency toward similar resource-allocation patterns.

But, and this is the more important implication, the short-run properties of the two

to profits per worker; and in our own country, profit sharing is not an exotic innovation but a current reality for many tens of millions of self-employed workers, professional partners, and people who work on commission or tips. (See my 1984 book, ch. 7.)

[15] An alternative approach to proving such propositions in a slightly different context is contained in my 1983 article. It is straightforward to generalize the present formulation to include capital, and relatively easy to verify that long-run properties are unaltered when the capital stock is treated as a choice variable. In long-run equilibrium, identical-twin wage and profit-sharing systems stimulate equal investment—to the point where the long-run marginal revenue product of capital equals the prevailing interest rate. What happens to capital formation out of long-run equilibrium can only be conjectured; but a fair guess might be that the relatively stable environment of a share economy—whose output is permanently maintained at the full capacity level—leads to an increased, steadier volume of investment over the business cycle.

systems (when pay parameters are quasi fixed) are quite strikingly different in the neighborhood of a long-run equilibrium position. From (58), (46), and (56), a profit-sharing system with a good-sized share component will be operating well inside the full-employment region $[\tau' > 0]$. (In long-run equilibrium, τ' is bounded below by 0, becoming ever larger as λ is bigger and as ω becomes smaller, approaching infinity as the pure wage component ω goes to zero and as λ approaches w^*L^*/P^*Y^*.) Even allowing for real world disturbances and *realpolitik* noncompetitive labor markets, a serious profit-sharing economy should remain at full employment. So it seems a fair generalization to say that, in the real world, a genuine profit-sharing system will be operating in the region $[\tau' > 0]$ whereas a wage system will be largely confined to the region $[\tau < 0]$. The wage variant of capitalism, unlike its profit-sharing cousin, cannot long be situated in a state of positive tautness because self-interested wage-economy firms will voluntarily bid up pay parameters.

There is then a marked difference in the degree of tension of the labor markets of wage and profit-sharing systems. A wage firm wants to hire as much labor as it is hiring under its current wage contract. But a profit-sharing firm wants to hire more labor than it is actually able to hire on the profit-maximizing contract parameters that it has itself selected.[16] The resolution of the seeming paradox is that while the profit-sharing firm desires more labor on the *old* contract, it will be made worse off if it tries to issue a *new* contract with higher pay parameters. (Indeed, this statement was demonstrated in the course of proving (58).)

It is important to note that it is not disequilibrium per se which causes unemployment, but rather a particular method of labor compensation (the wage system) in combination with disequilibrium. A profit-sharing system does *not* eliminate unemployment in a contractionary state by having such a high degree of pay flexibility that, in effect, wages are lowered to the point where long-run equilibrium is automatically maintained.[17] To see this point clearly, imagine a pair of "identical-twin" wage and profit-sharing economies, both in long-run stationary equilibrium with competitive labor markets, so that in both systems, worker pay equals the marginal revenue product of labor. Then subject the two systems to a contractionary shock and observe what happens in the short run.

In a profit-sharing economy, the marginal revenue product of labor, from (53) and (59), is

$$(66) \qquad R' = \gamma\beta M/\mu(Y^* - \alpha A),$$

while money pay (from Table 2) is

$$(67) \quad W = \beta M/(Y^* - \alpha A) \times\left[(1-\lambda)\frac{\omega(Y^* - \alpha A)}{\beta M} + \lambda\frac{Y^*}{L^*}\right].$$

Now whenever a profit-sharing economy is in long-run equilibrium, with (ω, λ) satisfying (58), then (66) and (67) must be equal, or $R' = W$. After a contractionary shock (say a decrease in A or M), it is straightforward to verify that money pay (67) declines by less than the marginal revenue product of labor (66) (provided $\omega > 0$). The marginal revenue product of labor will then be lower than pay, $R' < W$, yet all workers are retained by the firms. Thus, profit sharing does more than simply introduce some flexibility of wages. It builds in a permanent incentive for firms to want to retain their employees, not because of low pay, but because the marginal cost of an extra worker is less than the marginal revenue product created by that worker. In a wage system, on the other hand, firms always act to equate the marginal revenue product of labor with pay, and workers are consequently laid off after a contractionary shock.

Incidentally, it is straightforward to use the same identical-twin thought experiment to verify that not only is aggregate output

[16] This aspect is elaborated in my 1984 book.

[17] See my 1983 article for a more rigorous discussion.

and employment higher in a profit-sharing economy than a wage economy immediately after a contractionary shock to a long-run equilibrium state, but so is each employed worker's real pay. The conclusion about comparatively higher real pay in a share system holds as well for inflationary disturbances to a long-run equilibrium position, because there is at least some protection against higher prices.

Summing up, then, it seems a fair generalization to say that a serious profit-sharing economy will possess basically classical or monetarist macroeconomic properties very different from the short-run Keynesian underemployment characteristics of a wage economy. In a share economy, money is neutral and directly affects the price level, while having no effect on real aggregate economic variables. Resources are always fully utilized in a share system. The implication would appear to be that the central bank can directly and relatively easily control prices in a profit-sharing economy by regulating the supply of money, without having to worry about possibly adverse effects on employment and output.

VII. Wage and Profit-Sharing Economies Compared

It has been noted that a wage economy can plausibly be expected to function primarily in a regime where $\tau < 0$, whereas a profit-sharing economy should operate within the region $[\tau' > 0]$. The relevant conditions, I have argued, are

$$(68)\quad \alpha A + \beta M\gamma/\mu w < Y^* < \alpha A + \beta M\gamma/\mu\omega.$$

Throughout this section it is assumed that (68) describes the appropriate configuration of parameters, both initially and after unexpected displacements of the system.[18]

Table 3 compares the short-run macroeconomic properties of wage and profit-sharing systems in the regions where each is likely to be operating.

In order to be able to make meaningful comparisons between real pay, W/P, in both systems, some assumption of "comparability" must be made between pay parameters of wage and profit-sharing economies. The assumption made here is that real pay in the profit-sharing system should be the same as in the wage system—under the prices prevailing in the wage system, that is,

$$(69)\quad \gamma/\mu = (1-\lambda)\omega\gamma/w\mu + \lambda(Y^*/L^*).$$

(It is not difficult to verify that (69) is merely a rewriting of the long-run competitive labor market condition (58), which allows the fictional interpretation that both systems once upon a time started from the same initial equilibrium condition before being hit by the identical contractionary shock.)

Under conditions (68) and (69), from Table 3, output Y and real pay W/P are lower, while prices P are higher in a wage economy than in a comparable profit-sharing economy. This is the sense, then, in which conversion from a wage system to an equivalent-looking profit-sharing system yields unambiguously superior macroeconomic characteristics.

The basic short-run difference between sticky-pay-parameter wage and profit-sharing systems is no doubt exaggerated in my presentation, but it would, I feel, remain in modified form even after introducing additional real world frictions, inertias, and imperfections. Perhaps the contrast can be summed up as follows. In a wage system, prices are relatively rigid while quantities are relatively flexible and able to be influenced by demand management policies. In a share system, output prices are relatively flexible and under the control of monetary and fiscal policies, while quantities are relatively rigid at the full-employment level. Without relying on any fictitious "aggregate supply curve," which has little meaning in an imperfectly competitive environment where firms *set* prices so that there is always an excess supply of their products, the central theoretical result can nevertheless be conveniently stated

[18] The interested reader should be able to provide, from Tables 1 and 2, the correct analysis for those situations where (68) might not hold.

TABLE 3—MACROECONOMIC VARIABLES COMPARED IN THE TWO SYSTEMS

Variable	Wage Economy	Profit-Sharing Economy
Y	$\alpha A+\dfrac{\beta M\gamma}{\mu/w}$	Y^*
P	$\mu w/\gamma$	$\beta M/(Y^*-\alpha A)$
W/P	γ/μ	$\lambda\dfrac{Y^*}{L^*}+(1-\lambda)\omega\dfrac{Y^*-\alpha A}{\beta M}$

in the "as if" language of aggregate supply familiar to conventional macroeconomics.

A wage economy behaves in the short run as if aggregate supply were elastic at fixed prevailing prices (the as if Keynesian case). A profit-sharing economy behaves in the short run as if aggregate supply were inelastic at the full-employment level (the as if classical case).

Note that these statements describe the profit-maximizing Nash equilibrium behavior of a monopolistically competitive economy in the short run, when labor-payment contract parameters are fixed. The conclusions are not limited to the long run, or restricted to a perfectly competitive world. The share system thus behaves essentially like a classical macroeconomy, even while the classical preconditions are not being met. And the wage system, of course, behaves in the short run like the Keynesian macroeconomy that it is.

There is an interesting contrast, from Table 3, between the government's ability to influence prices and quantities in the two systems. Output in a profit-sharing economy automatically self-regulates at the full-employment level, independent of government policy or lack of policy. The world of Keynesian "underemployment equilibrium," on the other hand, with its possibility of (indeed its need for) using demand management to improve the level of aggregate output in the short run, with its attendant entourage of fiscal and monetary multipliers, rests crucially on the institutional assumption of a wage payment system. Change that particular labor payment feature to a profit-sharing arrangement and macroeconomic properties are dramatically altered for the better.

Compare the price equation of Table 3 for the two systems. In a wage economy, government policy has no *direct* effect on prices, which are determined strictly as a markup on costs. But in a share economy, the short-run price level is a direct function of aggregate fiscal and monetary variables and it does not depend upon short-run cost considerations. Government spending in a profit-sharing system crowds out private spending, and the aggregate effects show up only on the price level. Money is neutral in a share economy —monetary policy can be used powerfully and directly to determine the price level without affecting real economic activity. If there is an inflationary shock, say due to an increase in autonomous spending, the monetary authorities can hold the price level stable—without causing unemployment—merely by contracting the money supply. The share economy is a monetarist's dream—not just in long-run equilibrium, but in the short run with rigid labor contracts and monopolistic product markets.

A good litmus test for any market system is to observe how it reacts to changes in capacity. What happens if potential output Y^* is suddenly made larger, say because labor supply has unexpectedly increased?

A profit-sharing economy immediately raises its output level to the new capacity ceiling. Fresh labor is immediately absorbed and put to work producing additional goods and services, without having to wait for any long-run adjustment of pay parameters. From Table 3, the short-run effect of increased capacity on a profit-sharing economy is

greater output, lower prices, and higher real pay. The opposite conclusions hold when there is diminished potential to produce.

By contrast, in the wage system a firm is not interested in hiring additional workers on the existing labor contract. From Table 3, an increase in Y^* has no immediate effect on output, prices, or real pay for a wage system. Only if A, M, α, or β are increased, say through government policy, or if w is lowered, does a wage system absorb new entrants into the labor market.

The parameter μ is a measure of the degree of competitiveness of an economy. Higher values of μ mean that industry is less competitive. From Table 3, changes in μ have no short-term macroeconomic effects on a profit-sharing system, although there will be predictable long-term effects. By contrast, in a wage economy any industrial policy changing the degree of concentration will immediately move aggregate output, prices, and real pay in the expected direction, with macroeconomic performance being improved by increased competitiveness.

In the model of this paper, the coefficient γ stands for the marginal product of labor; its inverse, $1/\gamma$, measures the additional labor requirement per unit increment of output. If raw materials are employed in fixed proportions with output, an exogenous hike in the relative cost of materials could be given an interpretation within the model by appropriately increasing $1/\gamma$. Generally speaking, an adverse supply shock can be captured in the present framework by an autonomous deterioration of the marginal productivity parameter γ.[19]

[19] This is a standard trick, if somewhat heuristic. For some more details, see Rudiger Dornbusch and Stanley Fischer (1984, p. 410). Changes in sales taxes, employment subsidies, and the like can be given a similar interpretation. Note that I am assuming, for convenience, that a supply shock leaves the level of potential output Y^* unaltered. This may or may not be an appropriate assumption, depending on the context. The interested reader should be able to trace through, for example, what happens if γ and Y^* both change in the same proportion.

From Table 3, changes in γ have no short-term macroeconomic effects on a profit-sharing system. But a decline in the marginal productivity of labor has an immediate detrimental impact on output, prices, and real pay in a wage economy. The long-run effects of declining marginal productivity of labor are identical in both systems, involving basic adjustments in compensation parameters and real pay. But a share system allows such changes to come about gradually, through the competitive pressures of the market, without ever interrupting the smooth flow of full-employment output. A wage system, by contrast, responds to an adverse supply shock by an abrupt increase in unemployment and inflation that can be unsettling to society.

Wage capitalism is fundamentally a precariously balanced system. The slightest change—a momentary lowering of the desire to spend money on goods, say—can move it away from the razor-thin [$\tau = 0$] region where there is just full employment and pay is exactly competitive. A wage economy is at the mercy of any imbalances between γ, w, M, A, and the other variables or parameters of the system. A trifle more belligerence on the part of labor unions, a slight increase in the cost of imported raw materials, a bit less productivity than expected—may be enough to set off an explosive inflationary spiral, pushing up both prices and unemployment.

If productivity is less than anticipated, yet workers seek to maintain an inappropriately high level of real wages, even a very small discrepancy between labor's aspiration level and the profit-maximizing real wage

$$w/P = \gamma/\mu \tag{70}$$

may unleash an accelerating wage-price spiral, abetted by whatever indexation exists, that can ultimately be brought under control only by choking the economy, and the labor force, into submission through restrictive monetary and fiscal policies. When w is pushed up relative to γ, say because productivity has not increased as fast as expected, that just moves up prices in the same proportion, leaving the real wage intact. And unless

there is accommodating policy, unemployment results and output declines. Should the monetary authorities ratify the wage hike by increasing the money supply, inflation is created without dampening labor's underlying desire for an increased real wage.

A fundamental problem of the wage system is that prices are set by producers as a markup over wages and neither the government nor anyone else has a *direct* mechanism for changing the price level in the short run. From formula (43), P can only change as w, μ, or γ are altered. And there is no reason to expect a reliable or usable tendency for "the elasticity of demand, over the elasticity minus one, divided by the marginal product of labor" to vary systematically with business fluctuations.

So the only practical way to moderate prices in a wage economy is to moderate wage costs. Monetary or fiscal policies can slow down wage-push inflation only by throttling the economy into sufficiently low rates of employment to diminish money-wage demands: a very costly, indirect, inefficient, and inhumane way of controlling the price level, but the only one available under wage capitalism.

Table 3 displays an interesting contrast that may be relevant for issues concerning cost-push inflation. In a wage economy, the pay parameter w influences aggregate output and the price level, but not the real wage. In a share economy, it is the other way around —parameters ω and λ have no effect on output or prices, but do play a role in determining real pay. A cost-push money-wage increase in a wage economy lowers output and raises prices while leaving the real wage intact. But in a profit-sharing economy, any pushing up of pay parameters does nothing to aggregate output or prices, while it raises the level of real pay. If the parameters ω or λ are increased, that merely redistributes income in the short run from capital to labor without changing the overall size of the output pie.[20]

VIII. Conclusion

My own conclusion is that a profit-sharing economy has some natural tendencies toward sustained, noninflationary, market-oriented full employment. A profit-sharing economy can avoid dreaded Keynesian unemployment, even when conducting anti-inflationary monetarist policy. The wage variant of capitalism, on the other hand, does not have built-in stability and so must rely more heavily on skillful discretionary adjustments of financial aggregates in reacting to each unforeseen event as it occurs. Such questions as why wage capitalism is so prevalent and what can be done to change an economy from a wage system to a profit-sharing system must be left for another time.[21] But I hope it is clear from the analysis of this paper why an economy based on profit-sharing principles may conceivably offer some foundation for a permanent solution to the problem of stagflation.

[20] It might be thought, then, that there is a greater temptation for the median worker to attempt to push pay parameters above competitive levels in a profit-sharing economy than in a wage economy. Somewhat paradoxically, the exact opposite is true. See my 1984 book, ch. 8, for the details. It turns out that while it may be collectively rational for all workers together in a profit-sharing economy to push up pay parameters above competitive levels, it is not individually rational for a particular worker or union, who will not directly benefit, because on the margin the profit-sharing firm will automatically offset artificial pay-parameter increases by hiring more workers and driving down profits per worker, so pay remains at the level prevailing throughout the rest of the economy. In a wage system the opposite is true—it is individually rational for the median worker of a wage firm to push for higher wages no matter what workers in other firms are doing, but it is collectively irrational for the working class as a whole to push for higher wages.

[21] For some preliminary thoughts on these issues, see fn. 13. The welfare effects of changing from a sticky wage economy to a sticky share economy should be clear enough, even without a very sophisticated analysis. When outsider unemployed workers are effectively cut out of the wage economy, a significant slice of the national income pie evaporates—resulting in huge first-order Okun-gap losses of output and social welfare. A profit-sharing system stabilizes aggregate output at the largest possible national income pie, while permitting only small second-order Harberger-triangle losses to arise—for example, because a few crumbs have been randomly redistributed from workers in one firm to workers in another, or because the movement of resources in response to firm-specific shocks may be somewhat slowed.

REFERENCES

Benassy, Jean-Pascal, *The Economics of Market Disequilibrium*, New York: Academic Press, 1982.

Dixit, Avinash and Stiglitz, Joseph E., "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, June 1977, *67*, 297–308.

Dornbusch, Rudiger and Fischer, Stanley, *Macroeconomics*, 3rd ed., New York: McGraw-Hill, 1984.

Grandmont, Jean-Michel, *Money and Value*, Cambridge: Cambridge University Press, 1983.

Hart, Oliver, "Monopolistic Competition in the Spirit of Chamberlin: A General Model," *Review of Economic Studies*, 1985 forthcoming.

Hicks, John R., "Mr. Keynes and the 'Classics'; A Suggested Interpretation," *Econometrica*, April 1937, *5*, 147–59.

Malinvaud, Edmond, *The Theory of Unemployment Reconsidered*, Oxford: Basil Blackwell, 1978.

Meade, James E., *Stagflation*, Vol. 1: *Wage-Fixing*, London: Allen & Unwin, 1982.

______, "The Macroeconomic Implications of Monopolistic Competition," working paper, 1984.

Solow, Robert M., "Monopolistic Competition and the Multiplier," *Economic Journal*, 1985 forthcoming.

Varian, Hal R., *Microeconomic Analysis*, New York: W. W. Norton, 1984.

Weintraub, Sidney, *Our Stagflation Malaise*, Westport: Quorum Books, 1981.

Weitzman, Martin L., "Increasing Returns and the Foundations of Unemployment Theory," *Economic Journal*, December 1982, *92*, 787–804.

______, "Some Macroeconomic Implications of Alternative Compensation Systems," *Economic Journal*, December 1983, *93*, 763–83.

______, *The Share Economy*, Cambridge: Harvard University Press, 1984.

______, "Increasing Returns and the Foundations of Unemployment Theory: Reply," *Journal of Post Keynesian Economics*, Spring 1985, *7*, 403–09.

OPEC Behavior: A Test of Alternative Hypotheses

By JAMES M. GRIFFIN*

Just as the Great Depression characterized the 1930's, the Energy Crisis marked the 1970's. But what role will history ascribe to OPEC? Will it be considered a cartel, a loose confederation of producers who only ratified competitive prices, a group of producers seeking noneconomic goals, or whatever? While most students of energy economics label OPEC a cartel, increasingly this view has been challenged.[1] Recently, Paul MacAvoy (1982) has argued that the long-run trend of oil prices can be adequately explained by a competitive model. Proponents of the target revenue model (David Teece, 1982, and Jacques Cremer and D. Salehi-Isfahani, 1980) claim that rising oil prices and production cutbacks in the 1970's were not the result of collusive behavior, but rather each exporting country facing a backward-bending supply schedule, which results if exporting countries try to satisfy a target revenue for internal investment purposes. Walter Mead (1979) and A. D. Johany (1978) set forth a property rights explanation, arguing that the transfer of concession ownership from the international oil companies to the exporting countries in the 1970's lowered the effective discount rate leading to the observed price-output behavior. Finally, Theodore Moran (1982), a political scientist, rejects the conventional cartel interpretation in favor of an oil pricing model guided by Saudi Arabia utilizing its market power to maximize its security and influence rather than its wealth.

The standard practice to date has been to reach onto the shelf of economic models, to select one, to validate its choice by pointing to selected events not inconsistent with the model's predictions, and then to proceed with some normative exercise. For all practical purposes, the model is treated as a maintained hypothesis. Now, it is time to test these models since our interpretation of past and future oil price patterns depends critically on the model chosen. As a first step, the emphasis here is on testing the simplest version of the various models.

Section II outlines the determinants of an individual country's oil production level under the following four models: cartel, competitive, target revenue, and property rights. Section III compares the ability of each theoretical model to explain production levels in the various OPEC countries. Section IV considers the limitations of these results, contrasting them with non-OPEC oil producers. Section V recapitulates the principal findings.

I. Determinants of Production under Alternative Models

A. *Cartel Models*

The existing literature exhibits a variety of cartel models, each with unique empirical implications for an OPEC country's production decision. For example, Robert Pindyck (1978) has modeled OPEC as a monolithic wealth-maximizing monopolist solving an optimal control problem.[2] In contrast, Dermot Gately and John Kyle (1977) suggest OPEC follows a rule-of-thumb approach to pricing. Cartel models vary not only in their treatment of the nonrenewability of the re-

*Professor of Economics, Texas A&M University, College Station, TX 77843. I thank Steven N. Wiggins, Donald Deere, David Teece, the participants of the Yxtaholm Conference, and the referees for numerous helpful comments. Chris Colburn and Clifton Jones provided conscientious research assistance.

[1]For a literature review, see my study with David Teece (1982) and Dermot Gately (1984).

[2]Even though Pindyck (1982) now rejects the predictive value of his early approach, it is an important piece because it demonstrates that the monopoly price path will be initially much greater than the competitive price path.

source,[3] but also in the behavior of various OPEC countries.[4] The proliferation of models is not surprising, according to Morris Adelman (1982), because he contends OPEC's actual behavior wobbles between the dominant firm and market-sharing models depending on market conditions.

Following Adelman's suggestion, the empirical formulation of production is based on a market-sharing cartel model that allows for changing market shares over time. The simplest version of the market-sharing model posits constant market shares. The demand for OPEC oil (Q^O), is a derived demand, being the difference between world demand ($Q^W(P, A)$) and non-OPEC supply ($Q^{NO}(P, Z)$).

$$(1) \qquad Q^O = Q^W(P, A) - Q^{NO}(P, Z)$$

World demand depends on the real price of oil (P) and economic activity (A) while non-OPEC production depends on real price (P) and exogenous supply variables (Z). Given the cartel price (P), OPEC production is determined. Individual country production (Q_i) is some fraction (α_i^*) of total OPEC production (Q^O).[5]

$$(2) \qquad Q_i = \alpha_i^* Q^O \qquad i = 1, \ldots, n$$

To avoid simultaneity between Q_i and Q^O, equation (2) can be expressed as a fraction of other OPEC nations' production ($Q_i^{OO} = Q^O - Q_i$) as follows:

$$(3) \qquad Q_i = \alpha_i' Q_i^{OO} \qquad i = 1, \ldots, n$$

where $\alpha_i' = \alpha_i^*/(1 - \alpha_i^*)$.

Under a variety of cartel models market shares could vary with price, reserves, and other factors (see P. L. Eckbo). To allow for heterogeneous behavior, the market-share coefficient (α_i') is assumed to be a function of price:

$$(4) \qquad Q_i = \alpha_i^O P_t^{\gamma_i} Q_{it}^{OO} \qquad i = 1, \ldots, n; \quad t = 1, \ldots, T.$$

The empirical model to be estimated follows directly from equation (4):

$$(5) \qquad \ln Q_{it} = \alpha_i + \gamma_i \ln P_t + \beta_i \ln Q_{it}^{OO} + \varepsilon_{it} \qquad i = 1, \ldots, n; \quad t = 1, \ldots, T,$$

where ε_{it} is the standard error term. Tests for the following three variants of equation (5) were performed: constant market sharing ($\beta_i = 1$, $\gamma_i = 0$); market sharing ($\beta_i = 1$, $\gamma_i \gtrless 0$); and partial market sharing ($\beta_i > 0$, $\gamma_i \gtrless 0$). Partial market sharing suggests a looser cartel model since market-sharing considerations partially affect production decisions, but production cutbacks need not be proportional.

B. *The Competitive Model*

MacAvoy has added considerable credibility to the view that oil prices can be best explained by a model focusing on supply and demand rather than on cartel behavior. MacAvoy carefully avoids labeling his simulation model as competitive, choosing instead the terms "market fundamentals," "open market conditions," and "supply and demand." Nevertheless, the positively sloped supply functions for OPEC countries implicitly embody competitive behavior. MacAvoy primarily attributes price increases to supply disruptions unassociated with cartel ambitions: "Because of the supply interruptions prices would have risen to four-fifths or more of present prices under open market conditions... these conditions, and not OPEC, caused most of the crude oil price increases in the 1970's" (p. 57).

Besides MacAvoy's analysis, changing perceptions of user costs in the 1970's might well explain the large price increases. The 1960's and early 1970's experienced rapid demand growth and declining worldwide oil discovery rates which may have raised expected user costs.

The competitive model would specify current production (Q_{it}) as a function of price

[3] For examples, see Partha Dasgupta and Geoffrey Heal (1979) and Karim Pakravan (1984).

[4] See P. L. Eckbo (1976) for a description of the cartel core, the price pushers; and the expansionist fringe.

[5] For textbook examples, see my study with H. B. Steele (1980).

(P_t), perceived user costs (U_{it}), and current extraction costs (M_{it}):

$$(6)\quad Q_{it} = S_i(P_t, U_{it}, M_{it}) \qquad i = 1,\dots,n; \quad t = 1,\dots,T.$$

Unfortunately, data limitations do not allow estimation of equation (6) because user costs are unobservable and time-series data on extraction costs are not available.[6] Therefore, the empirical model is a simple log linear specification in which production is a positive function of the real oil price:

$$(7)\quad \ln Q_{it} = \alpha_i + \gamma_i \ln P_t + \varepsilon_{it} \qquad i = 1,\dots,n; \quad t = 1,\dots,T.$$

The underlying test is whether γ_i is positive, indicating a positively sloped supply schedule. It should be remarked that equation (7) is not devoid of cost and reserve phenomena as α_i and γ_i vary across producers, reflecting intercountry differences in cost and reserve conditions.

C. *The Target Revenue Model*

A popular, noncollusive explanation for oil pricing is the target revenue theory (Ali Ezzati, 1976; Cremer-Isfahani; and Teece). This theory argues that internal investment needs effectively determine "oil revenue needs." The former are constrained by the economy's ability to absorb investments and, for given prices, determine oil production. Once oil revenues satisfy the investment target, there is no incentive to produce more. Three critical assumptions underlie this model:[7] (*i*) a fixed and slowly expanding set of investment projects; (*ii*) the unacceptability of foreign investments; and (*iii*) oil revenues are the sole source of investment funds. The target revenue theory argues that production cutbacks occur in response to rising oil prices to equate oil revenues with investment needs. More formally, let I_{it}^* represent investment needs, the target revenue proponents postulate:

$$(8)\quad I_{it}^* = P_t Q_{it}.$$

By assumption, oil prices are exogenous to the producer as are investment needs, so we take logarithms and rearrange equation (8):

$$(9)\quad \ln Q_{it} = \alpha_i + \gamma_i \ln P_t + \delta_i \ln I_{it}^* + \varepsilon_{it} \qquad i = 1,\dots,n; \quad t = 1,\dots,T.$$

Increases in investment needs, *ceteris paribus*, result in proportionate ($\delta_i = 1$) increases in production.[8] For given investment needs, a price increase implies a proportionate output decrease ($\gamma_i = -1$). This constitutes the "strict version" of the target revenue theory. If OPEC countries are heavily influenced by target revenue considerations, but occasionally produce in excess of investment needs, I also test for a "partial version" of the theory ($\delta_i > 0$, $\gamma_i < 0$).

D. *The Property Rights Explanation*

Mead and Johany have adapted the property rights literature to provide yet another noncollusive explanation for world oil prices. Not only were the 1970's a time of rising oil prices, they also saw transfer in ownership and control of the oil concessions from the international oil companies to the producing countries. According to this theory, the transfer of ownership from the oil companies, who because of their impending loss of production were applying high rates of time discount, to the host countries, who tend to apply much lower discount rates, resulted in

[6] In the absence of producers' future oil price predictions, user costs are not observable. Lease bonus prices in competitive auctions may form the basis for approximations to future price expectations, but are themselves limited to countries with (a) competitive leasing and (b) reasonably homogeneous prospects. For an attempt to estimate user costs in OPEC countries, see Pakravan.

[7] See Teece, pp. 66–67.

[8] Since investment needs are not directly observable, I use as a proxy real fixed domestic capital formation, converted to US $ and lagged one quarter, as reported in the IMF *International Financial Statistics*.

sharp production cutbacks and rising prices. In a Hotelling model with zero extraction costs, the real price of oil will rise at the real discount rate. The switch from a high real discount rate to a low real rate leads to sharply higher prices and lower current production (see Dasgupta and Heal). After full adjustment to the lower discount rate, prices rise at the new real rate of discount. Under the property rights explanation, production will be primarily influenced by the percentage of government controlled production (G_{it}); increases in the percentage of government controlled production result in production cutbacks ($\varnothing_i < 0$).[9]

$$\ln Q_{it} = \alpha_i + \varnothing_i G_{it} + \varepsilon_{it} \qquad i=1,\ldots,n; \quad t=1,\ldots,T. \tag{10}$$

II. Empirical Tests

This section examines the empirical evidence supporting these four hypotheses. With the exception of the property rights model, quarterly data for the period 1971:I through 1983:III were utilized. The sample period is particularly rich as it includes a substantial period prior to the October 1973 Arab-Israeli war and the quadrupling of crude prices. Likewise, the data through 1983:III contains observations including the decline in crude prices from \$34 to \$29 in the spring of 1983. Quarterly, as opposed to annual, data allow a much closer examination of short-run production adjustments in response to price and other factors. Due to data deficiencies for Ecuador and Gabon, only 11 of the 13 OPEC countries are examined. The sample is truncated for Iran to delete the period following the revolution in Iran (1978:III) and for Iraq after the start of the Iran-Iraq war (1980:III). Price and production data are taken from *Petroleum Intelligence Weekly* and the *U.S. Monthly Energy Review*.[10]

A. The Cartel Model

Table 1 reports the relevant *OLS* regression estimates for the four models and Table 2 shows the results of hypothesis tests. The latter are based on *t*-tests for individual parameters and for joint hypotheses, an *F*-test is applied. Beginning with Table 1, Part A, note that the coefficient for other OPEC production (Q_{it}^{OO}) is statistically significant and positive in 10 of the 11 OPEC countries. Only Iraq, long noted for independent behavior, exhibits a negative but insignificant coefficient. Also, the cartel model exhibits frequent statistically significant price terms, indicating that market shares change with price. Consequently, the constant market-sharing model is decisively rejected (see Table 2). On the other hand, the market-sharing model ($\beta_i = 1,\ \gamma_i \gtrless 0$) cannot be rejected for 5 of the 11 countries including Qatar, United Arab Emirates (U.A.E.), Libya, Iran, and Nigeria.[11] Even Saudi Arabia appears to vary production in concert with other OPEC countries. This result contradicts dominant firm models where Saudi Arabia would act as the market leader and vary production inversely to the competitive output levels of other producers, including the rest of OPEC. Based on this evidence, OPEC appears to be a real cartel with at least partially effective output coordination. Although it was not possible to test Moran's security maximization model directly, these results indicate that OPEC is much more than Saudi Arabia.

The coefficients on price suggest some interesting behavioral implications about market shares. Kuwait, Qatar, Libya, and Venezuela appear willing to accept lower market shares in a high price regime. For Libya and Venezuela, their preferences for higher prices

[9]As a measure of the percentage of government controlled oil production, I use the percent of government equity oil as reported in the OPEC Annual Statistical Bulletin. While seemingly a perfect measure, government equity oil is not synonymous with government controlled production. The decision to produce and sell government equity oil is not necessarily made by the government as the remaining concessionaires may make the production decision and simply compensate the government. There is, however, no better measure available.

[10]A copy of the data appendix is available on request.

[11]Additionally, the models were corrected for first-order autocorrelation and the conclusions were not found to be affected. Results are available on request.

TABLE 1—ESTIMATES OF ALTERNATIVE MODELS APPLIED TO 11 OPEC COUNTRIES[a]

	A. Cartel			B. Competitive		C. Target Revenue			D. Property Rights	
Country	$\ln Q_{it}^{OO}$ (β_i)	$\ln P_t$ (γ_i)	$\bar{R}^2$	$\ln P_t$ (γ_i)	$\bar{R}^2$	$\ln I_{it}^*$ (δ_i)	$\ln P_t$ (γ_i)	$\bar{R}^2$	G_{it} ($\varnothing_i$)	$\bar{R}^2$
Saudi Arabia	.74	.29	.49	.15	.17	−.004	.14	.12	.61	.74
	(.13)	(.04)		(.04)		(.07)	(.11)		(.11)	
Kuwait	1.41	−.39	.90	−.48	.59	−.15	−.39	.59	−.61	.51
	(.12)	(.03)		(.05)		(.10)	(.09)		(.18)	
Qatar	.88	−.03	.73	−.11	.14	N.A.	N.A.		−.03	−.09
	(.08)	(.02)		(.04)					(.03)	
U.A.E.	1.00	.20	.80	.09	.09	.11	−.06	.11	.66	.74
	(.08)	(.02)		(.04)		(.06)	(.08)		(.12)	
Iraq	−.06	.29	.49	.30	.50	.27	−.02	.67	.55	.46
	(.48)	(.05)		(.05)		(.06)	(.08)		(.20)	
Libya	.72	−.27	.59	−.33	.45	−.38	−.06	.45	−.57	.27
	(.18)	(.05)		(.05)		(.22)	(.16)		(.26)	
Algeria	.74	.005	.69	−.07	.07	.34	−.28	.22	.39	−.05
	(.07)	(.02)		(.03)		(.10)	(.07)		(.55)	
Iran	.88	.05	.69	.09	.32	−.004	.08	.25	.21	.75
	(.15)	(.02)		(.02)		(.05)	(.04)		(.05)	
Nigeria	1.13	.10	.68	−.01	−.02	−.05	−.008	−.03	.18	.001
	(.11)	(.03)		(.05)		(.11)	(.07)		(.18)	
Indonesia	.56	.26	.83	.19	.54	.14	.08	.52	1.49	.87
	(.06)	(.02)		(.03)		(.09)	(.07)		(.18)	
Venezuela	.17	−.25	.75	−.26	.73	−.25	−.15	.78	−.33	.68
	(.08)	(.02)		(.02)		(.06)	(.03)		(.07)	

[a]Standard errors are shown in parentheses.

TABLE 2—SUMMARY OF HYPOTHESIS TESTS

	A. OPEC Countries		B. Non-OPEC Countries	
Models	Do Not Reject	Reject	Do Not Reject	Reject
1. Cartel				
Constant Market Sharing	1	10	0	11
Market Sharing	5	6	2	9
Partial Market Sharing	11	0	8	3
2. Competitive	6	5	10	1
3. Target Revenue				
Strict Variant	0	10		
Partial Variant	9	1		
4. Property Rights	6	5		

match their traditional positions at OPEC meetings. Of those countries found to increase market share in response to rising prices, we find a curious mixture ranging from Saudi Arabia to Indonesia. The results for Saudi Arabia seem consistent with a willingness to use production increases to attenuate price increases whereas a positive coefficient for Indonesia and Nigeria reflect some tendency to increase production at higher prices as would be expected from small producers with competitive fringe tendencies. In sum, the evidence favoring a partial market-sharing model is impressive.[12]

[12]Even the addition of lagged oil reserves to equation (5) does not alter this basic conclusion, as its coefficient

B. *The Competitive Model*

The results for the competitive model in Table 1 reveal generally anomalous results. The hypothesis of a positive coefficient ($\gamma_i > 0$) on price is rejected in 5 of the 11 countries. Significantly negative estimates are observed for Kuwait (−.48), Qatar (−.11), Libya (−.33), Algeria (−.07), and Venezuela (−.26).[13] Furthermore, comparison of the $\bar{R}^2$ for both the cartel and the competitive models reveals a sharp dichotomy in explanatory power. Additionally, it is possible to test the competitive model as a nested case of the cartel model. Even for the 6 countries in Table 2 for which the competitive model could not be rejected, the competitive restriction is rejected for all but Iraq. Attempts to improve the performance of the competitive model by the introduction of a lagged oil reserves variable in equation (7) failed to alter this conclusion.[14] Even experimentation with a dynamic demand formulation, allowing for lagged price effects likewise failed to alter the overall conclusion.[15]

C. *The Target Revenue Model*

Of the 10 OPEC countries for which investment data is available, all 10 are found to reject the strict formulation of the target revenue theory ($\delta_i = 1$, $\gamma_i = -1$). On the other hand, these factors could be influencing production decisions through a partial version of the target revenue model ($\delta_i > 0$, $\gamma_i < 0$). Table 1, Part C, shows there are only 3 countries for which investment is statistically significant and positive, and only one country (Algeria) with both variables having correct, statistically significant coefficients. Interestingly, Algeria is the country Cremer and Isfahani cite as an example of their theory. The null hypothesis test in Table 2 (H_O: $\gamma_i < 0$, $\delta_i > 0$) reveals that the partial version of the model, unlike its strict counterpart, is difficult to reject statistically despite the lack of evidence to support the theory.[16] Attempts to utilize a trended investment series to better approximate investment "needs" failed to improve the model's performance.

D. *The Property Rights Model*

The data set for this test utilized annual observations for the period 1971 to 1981, the last year for which annual data on the percent of government equity oil production (G_{it}) is available. Table 2, Part B, reveals a statistically significant and negative coefficient for 3 of the 11 countries (Kuwait, Libya, and Venezuela). Nevertheless, as indicated in Table 2, the null hypothesis (H_O: $\varnothing_i < 0$) could not be rejected for 6 of the 11 countries. The mixed performance of the property rights model is not unexpected since, for example, in 1972 Iraq nationalized the Iraq Petroleum Company and expanded capacity from 1.7 to 4.0 million barrels per day. In sum, this simple test of the property rights model provides little support for this hypothesis.

III. Caveats and Extensions

The preceeding results give considerable support to the cartel model. Not only does this model apply to a greater number of countries, but the overall level of explanation dominates other models. But to what degree are these results dependent on the test devised for each model? Perhaps due to spurious correlations, any randomly selected country's observations might fail to reject the cartel model. Alternatively, the power of the

is positive and statistically significant in only 3 countries (and none under the autoregressive model). In such cases, the conclusions concerning β_i are unaffected.

[13] With the autoregressive model, price tended to be positive and statistically significant in only four cases.

[14] Here the reserves coefficient was positive and statistically significant in 5 countries, causing a sign change in γ_i for U.A.E. from .09 to −.14, strengthening the conclusion a bit.

[15] In the dynamic formulation, γ_i was not positive and significant for any of the OPEC countries.

[16] Estimation using the autoregressive model found only Iraq to have a positive, statistically significant coefficient on I_{it}^*(.28), and no country showed both variables having the signs suggested by the target revenue theory.

TABLE 3—ESTIMATES FOR CARTEL AND COMPETITIVE MODELS APPLIED TO 11 NON-OPEC COUNTRIES[a]

	A. Cartel			B. Competitive	
Country	$\ln Q^O_{it}$ (β_i)	$\ln P_{it}$ (γ_i)	$\bar{R}^2$	$\ln P_{it}$ (γ_i)	$\bar{R}^2$
Argentina	−.17	.04	.19	.05	.16
	(.15)	(.03)		(.03)	
Brunei/Malaysia	.32	.32	.59	.30	.60
	(.37)	(.08)		(.07)	
Canada	.32	−.04	.20	−.06	.08
	(.20)	(.04)		(.04)	
China	.72	.62	.89	.57	.85
	(.31)	(.06)		(.07)	
Egypt	−1.16	.34	.32	.41	.28
	(.89)	(.18)		(.18)	
India	−.05	.22	.66	.28	.46
	(.36)	(.07)		(.09)	
Mexico	−1.29	.67	.83	.75	.75
	(.54)	(.11)		(.13)	
Norway	.75	1.73	.70	1.68	.72
	(1.66)	(.34)		(.31)	
U.K.	−.61	3.32	.61	3.36	.65
	(3.98)	(.81)		(.73)	
USSR	−.34	.19	.79	.19	.82
	(.15)	(.03)		(.03)	
U.S.	−.14	−.06	.62	−.05	.50
	(.07)	(.01)		(.02)	

[a]Standard errors are shown in parentheses.

cartel model may be sufficiently weak that most countries would fail to reject it. This section examines such questions in order to assess the confidence one can attach to the preceeding results.

A. *Non-OPEC Producers*

To test the conjecture that any oil producer would fail to reject the cartel model, I apply both the competitive and cartel models to non-OPEC producers with daily production rates in excess of 300,000 barrels in the 1971–82 period. The data set consists of 11 countries, which include Argentina, Brunei/Malaysia, Canada, China, Egypt, India, Mexico, Norway, United Kingdom (U.K.), USSR, and the United States (U.S.). Annual data (1971–82) were utilized in place of quarterly data due to limited data availability. One would expect, *ex ante*, the behavior of these countries to reject the cartel model and to not reject the competitive model. Table 3 presents the parameter estimates of the two competing hypotheses for the non-OPEC producers and Table 2, Part B, reports the hypothesis test results. Table 2, Part B, reveals that the power of the partial market-sharing hypothesis is not strong as for 8 of the 11 non-OPEC countries cartel behavior could not be rejected. These results can be directly traced to the fact that H_O: $\beta_i > 0$ is virtually equivalent to H'_O: $\beta_i = 0$. Consequently, it is noteworthy that the cartel model is rejected for 3 countries. On the other hand, the competitive model is rejected for only one country, the U.S., which was subject to price controls over much of the sample period.

A closer comparison of the cartel and competitive models for these non-OPEC countries reveals some striking differences. Whereas for the 11 OPEC countries, the coefficient on OPEC output, β_i, was positive and statistically significant for 10 of the 11 countries, only in one (China) of the 11

non-OPEC countries was this true.[17] Additionally, whereas for OPEC countries, the explanatory power of the cartel model substantially exceeds that of the competitive model, this is not the case for non-OPEC countries.

B. *Divergent, Disequilibrium Price Expectations*

The competitive label might be resurrected for the OPEC countries by appeal to a Hotelling story linking rising prices to current production cutbacks. As the discussion of the competitive model indicated, if changes in perceived user costs are correlated with price changes, the competitive model's price coefficient may be biased downward. Suppose OPEC producers altered their perceived user costs during the early 1970's in anticipation of future shortages. They would then have cut production, but the effect may have been incorrectly attributed to price, since user costs are omitted in equation (5).

There are several problems with such an explanation. First, one must wonder why the non-OPEC suppliers were not appreciably influenced by such a change in user cost perceptions since they rather uniformly exhibit a positive relationship between price and production in Table 3. Second, even if the expectations of the OPEC nations diverged from the non-OPEC producers, such expectations seem likely to be confined to the period 1971–74, since, thereafter, prices had effectively quadrupled. Thus, if the negative relationship between price and production is due to divergent expectations, one would expect to observe a positive price/quantity relationship for the period excluding these possibilities (1975–83). Tests of the competitive model for the period beginning 1975:I are quite revealing. Instead of showing a greater preponderance of positive price coefficients, as implied by disequilibrium price expectations, the competitive model was rejected for all countries except Iraq. Hence, the positive price coefficients in Table 1 which are supportive of the competitive model originated primarily from the period 1971–74. Such a result is quite plausible if cartel behavior only began with the 1973 Oil Embargo. In sum, divergent expectations of user costs do not appear capable of salvaging the competitive model.[18]

C. *Attempts to Group OPEC Countries*

With the recognition that OPEC cannot be treated in monolithic terms, researchers have found it convenient to group OPEC countries by behavioral characteristics. For example, Eckbo postulated three groups—the "price pushers" (Iran, Venezuela, Algeria, and Gabon), the "expansionist fringe" (Indonesia, Nigeria, Iraq, and Ecuador), and the "cartel core" (Saudi Arabia, Kuwait, U.A.E., Qatar, and Libya). In a statistical sense, is it meaningful to form such groups? Tests on Eckbo's groupings, excluding Ecuador and Gabon, were performed using the cartel model and allowing for differential country intercepts to see if common values for β_i and γ_i could be maintained within each of these three categories. In each group, the hypothesis of similar responses was decisively rejected.[19] While behavioral groupings are superior to a monolithic view of OPEC, such groupings must be viewed as rough approximations, useful for pedagogical purposes.

IV. Summary and Implications

The scope of this paper—to test these alternative theories of OPEC—is necessarily ambitious and controversial. Nevertheless, resolution of these differences is critical, not only for historical purposes, but also for the methodology of economics and its implications for public policy. Methodologically,

[17] Canada's coefficient β_i becomes marginally significant when correction for autocorrelation is made.

[18] For a similar conclusion, see Pakravan.

[19] For the price pushers, $F = 39.6$ with $F^*_{.01}(4,124) = 3.47$; for the expansionist fringe, $F = 17.5$ with $F^*_{.01}(4,132) = 3.47$; and for the cartel core, $F = 53.2$ with $F^*_{.01}(8,240) = 2.65$.

tests of these competing theories should produce clear differences unless, of course, monopolistic and competitive models yield scientifically indistinguishable differences. The approach has been to formulate simple models reflecting the key determinants of production behavior under alternative OPEC theories. These models are inherently simplifications of each theory in a variety of ways. Clearly, the proponents of each theory could postulate more elegant and complete empirical specifications of their positions. While the results here must be viewed as a first step, it appears doubtful that more elegant specifications will overturn the principal findings here.[20]

Perhaps the most striking aspect of the empirical tests is the clear-cut nature of the results. First, among OPEC countries, the partial market-sharing cartel model could not be rejected for all 11 countries, whereas frequent rejections are observed for the other theories. Second, in terms of the ability of the various models to explain production, the partial market-sharing cartel model dominates the competitive model. Third, in comparisons with 11 non-OPEC countries we observe the opposite tendency—the competitive model could not be rejected for 10 of the 11 non-OPEC producers.

This prompts the all important unanswered question: "If OPEC is indeed a cartel, why have Friedman's predictions proven so wide of the mark?" Traditionally, economists have viewed cartels as fragile entities with limited power to raise price appreciably above competitive levels and, if successful for a time, unable to sustain such increases over the longer run. Why do OPEC countries ostensibly follow some form of market-sharing behavior, especially since many of them have very small market shares and large incentives to cheat? Hopefully, future research will provide a richer cartel model, better capable of explaining the surprising stability of OPEC.

[20] Interestingly, C. W. Hope and P. H. Gaskell (1984) reach a similar conclusion using an optimal control approach.

REFERENCES

Adelman, Morris A., "OPEC as a Cartel," in J. M. Griffin and D. J. Teece, eds., *OPEC Behavior and World Oil Prices*, London: Allen & Unwin, 1982.

Cremer, Jacque and Salehi-Isfahani, D., "A Competitive Theory of the Oil Market: What Does OPEC Really Do?," Caress Working Paper, No. 80–4, mimeo., University of Pennsylvania, 1980.

Dasgupta, Partha S. and Heal, Geoffrey M., *Economic Theory and Exhaustible Resources*, Cambridge: Cambridge University Press, 1979.

Eckbo, P. L., *The Future of World Oil*, Cambridge: Ballinger, 1976.

Ezzati, Ali, "Future OPEC Price and Production Strategies as Affected by its Capacity to Absorb Oil Revenues," *European Economic Review*, August 1976, *8*, 107–38.

Friedman, Milton, *Newsweek*, March 4, 1974 and September 15, 1980.

Gately, Dermot, "A Ten-Year Retrospective: OPEC and the World Oil Market," *Journal of Economic Literature*, September 1984, *22*, 1100–14.

______ **and Kyle, John**, "Strategies for OPEC's Decisions," *European Economic Review*, November 1977, *10*, 209–30.

Griffin, J. M. and Steele, H. B., *Energy Economics and Policy*, New York: Academic Press, 1980.

______ **and Teece, David J.**, *OPEC Behavior and World Prices*, London: Allen & Unwin, 1982.

Hope, C. W. and Gaskell, P. H., "The Oil Price Without OPEC: Back to the $3 Barrel," paper presented at the 1984 BIEE/IAEE Conference, Cambridge, U.K., April 1984.

Johany, A. D., "OPEC is Not a Cartel: A Property Rights Explanation of the Rise in Crude Oil Prices," unpublished doctoral dissertation, University of California-Santa Barbara, 1978.

MacAvoy, Paul, *Crude Oil Prices as Determined by OPEC and Market Fundamentals*, Cambridge: Ballinger, 1982.

Mead, Walter J., "The Performance of Government Energy Regulations," *American Economic Review Proceedings*, May 1979,

69, 352–56.

Moran, Theodore, "Modeling OPEC Behavior: Economic and Political Alternatives," in J. M. Griffin and D. J. Teece, eds., *OPEC Behavior and World Oil Prices*, London: Allen & Unwin, 1982.

Pakravan, Karim, "Estimation of User's Cost for a Depletable Resource such as Oil," *Energy Economics*, January 1984, *6*, 35–40.

Pindyck, Robert S., "Gains to Producers from the Cartelization of Exhaustible Resources," *Review of Economics and Statistics*, May 1978, *60*, 238–251.

______, "OPEC Oil Pricing, and the Implications for Consumers and Producers," in J. M. Griffin and D. J. Teece, eds., *OPEC Behavior and World Oil Prices*, London: Allen & Unwin, 1982.

Teece, David, "OPEC Behavior: An Alternative View" in J. M. Griffin and his *OPEC Behavior and World Oil Prices*, London: Allen & Unwin, 1982.

Further Results on Inventories and Price Stickiness

By RICHARD A. ASHLEY AND DANIEL ORR*

The objectives of this paper are first, to sharpen a microeconomic definition of price stickiness and to offer criteria for its presence and indications of its potential magnitude; second, to distinguish two motives for holding inventories: production smoothing and uncertainty buffering: and to show how price stickiness may accompany the first of these in the monopoly firm; and third, to develop a model in which inventory backorders play a minimally significant part. Our efforts are in part an elaboration or extension of Alan Blinder's 1982 paper, which initiated the "inventory control and price stickiness" discussion, but which focused on uncertainty buffering. As in that paper, substantive attention here is confined to the case of a monopoly supplier.

By focusing on monopoly behavior exclusively, we are ignoring questions of social optimality, since the competitive framework requisite for comparison is missing. But despite the neglect for now of those welfare-related questions, there are interesting issues to be found entirely within the domain of monopoly. As in Blinder's paper, the central question is: does an intertemporally optimizing monopolist generate a price sequence that looks "sticky?" Our shift in focus from uncertainty buffering to production smoothing adds, we think, substantially to the interest of that question, and to the ability to answer it interestingly.

When are a monopolist's prices sticky, and when are they responsively flexible? That is not a trivial question. Much of the discussion of price stickiness (and the related issue of administered prices: see Gardner Means, 1975) seems to be based on a view that if shocks or shifts in demand or cost occur more frequently than changes in price, or if the price changes that are made in response to shocks are smaller than comparative static analysis indicates they should be, then prices are sticky. The implicit conclusion is that the monopolist fixes price in order to exploit the market, or to deter entry.

That outlook, as we will show, overestimates the optimal degree of price responsiveness by an inventory-holding monopoly. A firm's optimal prices and outputs cannot be described as a sequence of independent one-period static solutions, because interperiod dependencies arise from current production for future sale, a pattern that is created by carrying inventory. Price behavior when time periods are linked should not be expected to resemble that which would be generated in a sequence of independent static equilibria.

The effects of temporal interdependency on price movement are worth analyzing because the analysis gives a clearer idea of the patterns which are a natural part of optimizing behavior in the monopoly firm. Blinder's contribution was to show that prices move less than comparative static analysis predicts in the uncertainty buffering monopoly. But they do move, and in predictable directions. What we will establish here is that inventory management for production smoothing can lead to pricing behavior that contains genuine surprises, behavior far more in keeping with a reasonable meaning of "price stickiness" as the term is used in macroeconomic discussion. The price stickiness that we will look for is here given as

Definition 1: A price is *sticky* if it fails to adjust at all despite changes in the predicted values of cost or demand parameters; or if cost or demand shocks, whether foreseen or unforeseen, appear to elicit no price response; or if price moves up when comparative static analysis indicates it should move down (or vice versa.)

*Assistant Professor and Professor, respectively, Department of Economics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. The useful comments of Douglas McTaggart, Warren Weber, and the referee are acknowledged.

Definition 1 focuses on the presence and direction of price changes; by contrast, Blinder's contribution was to observe the diminished magnitudes of price changes when inventories are used to buffer stochastic shocks; his discovery might better be described as "price sluggishness."

In relying upon the certainty-equivalent model pioneered by Charles Holt et al. (1958), Blinder was forced to accept two consequences that we seek to avoid: 1) his model implies that inventories will frequently be negative, that is, the firm will persistently be backordered; and 2) the focus is upon price, output, and inventory responses to purely random demand fluctuations or to a stationary [$AR(1)$] stochastic demand process; the firm in that model cannot readily form optimal plans to meet seasonal demand peaks or other forecasted nonstationary demand changes. Our model, by contrast, will be free of those weaknesses, but in turn is unable to buffer in the presence of stochastic demand.

The precise sense in which we use the terms "buffering" and "smoothing" is given by

Definition 2: To "buffer" is to accommodate or offset movements in demand that are produced by a stationary stochastic process.

Definition 3: To "smooth" is to accommodate or offset movements in demand which are forecast with certainty, and which may be nonstationary.

Our first major result, to be developed in Section II, is that actual price stickiness per Definition 1 is a possible consequence when price, output, and inventory are managed for optimal production smoothing in a monopoly firm. We work in a dynamic, full-certainty model which has had a venerable place in the literature but little recent use. In Section I we use that model to develop results on planning horizons, which undergird the subsequent analysis of pricing behavior. The exposition of Section I not only provides necessary background, it also reveals the structure of the production smoothing process.

I. Inventories, Horizons, and Production Smoothing

Our model goes back to the late 1930's, when Edwin S. Shaw (1940) analyzed the effect of inventory on output adjustment when demand increases through time. Since then, the model has found a range of uses, most conspicuously in the major paper of Franco Modigliani and Franz Hohn (1955) and in Kenneth Arrow, Samuel Karlin, and Herbert Scarf (1958; for example, chs. 4 and 5). In those earlier applications the model was used to determine a firm's optimal production response when demand is nonstationary, or when the fluctuation pattern is dominated by a deterministic seasonal or cyclical component. Those earlier applications treated demand as a sequence of quantity requirements through time, and price did not enter as a decision variable. In our version, price and output are both explicit instruments of the firm.[1]

One major result will be exploited heavily in the following: except in some instances when demand is monotonically increasing through all future periods, an endogenously determined *optimal horizon* will exist, and the optimal price-output-inventory policy can be found up to that horizon without reference to any demand or other information concerning periods that lie beyond it. (This result was first published by Modigliani-Hohn, and elaborated by Abraham Charnes, Jacques Drèze, and Merton Miller, 1966; Dwight Lee and Orr, 1977; and others.) In this section, the conditions which locate an optimal horizon will be described, and an optimal price-output-inventory plan will be qualitatively characterized up to the optimal horizon. From that characterization we will see how a planned sequence of prices adjusts optimally to a forecast pattern of demand; and also how the optimal price sequence adjusts to a demand shock, defined as a change in a demand parameter within the optimal horizon, but which is detected only after the price-output program is under way.

[1]The results of this section are developed differently by Dwight Lee (1972).

Our analysis can incorporate foreseeable variations in both cost and demand, but for simplicity we will treat cost as constant through time.

The precise nature of an optimal horizon is brought out in these definitions:

Definition 4: A *forecast horizon*, denoted F, is the most distant future period through which the firm has demand forecast information in reliable deterministic form.

Definition 4 depicts a firm that knows the demand function for its output over all consecutive future periods, up to and including period F. Hereinafter, F is finite.

With F as the forecast horizon, and with x_t and s_t the values of output and sales in period t, our initial statement of the firm's objective is

$$(1)\quad \underset{\substack{x \geq 0 \\ s \geq 0}}{\text{Max}}\, \pi = \sum_{t=1}^{F} \theta^{t-1}\left[R_t(s_t) - C(x_t) - h \cdot I_t\right],$$

subject to the constraints

$$(2)\quad I_t = I_0 + \sum_{j=1}^{t}(x_j - s_j) \geq 0, \qquad t = 1, \ldots, F-1$$

$$(3)\quad I_F = I_0 + \sum_{j=1}^{F}(x_j - s_j) = 0,$$

$$R_t(s_t) = s_t \cdot D_t(s_t); \quad d^2R_t/ds_t^2 < 0,$$

where I_t is inventory[2] on hand at the end of period t; D_t is the demand relation in period t (which shifts from period to period in a known way); R_t is sales revenue; C is the cost of producing goods for placement in inventory (with $C' > 0$, $C'' > 0$); $h \cdot I_t$ is the cost of maintaining inventory throughout period t; and θ is the discount factor. We select a linear relationship between inventory levels and inventory cost for two reasons: because it has been a widely used—indeed, a standard—representation in the literature of inventory theory;[3] and because it has been found by Blinder (p. 347) to cause strange output responses in the linear-quadratic firm of his certainty-equivalent stochastic model.

The nonnegativity constraints on inventory are crucial. Without them, the firm could embark on a program of unlimited back-order accumulation, selling goods today with the promise of future delivery, and with unlimited deferral of production.

In contrast to the nonnegativity constraints imposed here, Blinder's formulation explicitly allows negative inventories. In fact, David Schutte notes that Blinder's model is better viewed "as a model of optimal order backlogs" (1983, p. 815), than as a model of optimal inventories since the stationary level of inventories in his model is negative whenever production is positive.

The zero-inventory constraint at the end of the forecast horizon ($I_F = 0$) is innocuous. Some specific value of I_F must be selected to close the solution for an optimal $x_1, \ldots, x_F$ and $s_1, \ldots, s_F$. But any nonnegative value could be chosen.[4]

With that apparatus in place, we proceed to

Definition 5: An *optimal program* over the horizon F is the sequence of sales and output decisions that solves the constrained maximization problem ((1)–(3)):

$$D^*(F) = (s_1^*, x_1^*, \ldots, s_F^*, x_F^*).$$

[2] Throughout our analysis, as in most of the vast existing literature on inventory theory, the term inventory refers to holdings of finished goods. Input materials and work in process usually are ignored.

[3] Useful brief discussions of the costs of holding inventory are found in Holt et al. and in George Hadley and Thomson Whitin (1963).

[4] The optimal horizon might be displaced in time if a different condition, for example, $I_F = Q > 0$, were imposed. $I_F = 0$ is natural, however, barring special reasons for choosing a different value.

Definition 6: A *truncated optimal program*, with truncation at K,

$$D_K^*(F)=(s_1^*,x_1^*,\ldots,s_K^*,x_K^*),$$

is the first K periods of the F-period optimal solution, $K<S$.

We are now ready for the punchline:

Definition 7: An *optimal horizon* occurs in period T if and only if

$$D_T^*(F)=D^*(T)$$

and
$$D_{T-K}^*(F)\neq D^*(T-K),$$
$$K=1,\ldots,T-1.$$

Thus, an optimal program through period T, which disregards available forecast information for periods $T+1$ through F, coincides with the T-period truncation of the optimal program over the forecast horizon; and if T is an optimal horizon, that coincidence occurs in no earlier period. When T has been located, information for periods beyond T is in a sense superfluous.

It remains to state

LEMMA: *In an optimization problem of the form* ((1)–(3)), *there will always exist an optimal horizon* $T\leq F$.

PROOF:

By familiar Kuhn-Tucker procedure, the Lagrangian

$$L(x,s,\lambda,\mu)=\pi+\sum_{t=1}^{F-1}\lambda_t\left[I_0+\sum_{j=1}^{t}(x_j-s_j)\right]$$
$$+\mu\left[I_0+\sum_{t=}^{F}(x_t-s_t)\right]$$

yields the first-order conditions

$$L_{x_t}=-\theta^{t-1}C'(x_t)-h\sum_{j=t-1}^{F-1}\theta^j$$
$$+\sum_{j=t}^{F-1}\lambda_j+\mu=0,\quad x_t\geq 0,$$

$$or\quad L_{x_t}<0,\quad x_t=0,\qquad t=1,\ldots,F;$$

$$L_{s_t}=\theta^{t-1}R_t'(s_t)+h\sum_{j=t-1}^{F-1}\theta^j$$
$$-\sum_{j=t}^{F-1}\lambda_j-\mu=0,\qquad s_t\geq 0,$$

$$or\quad L_{s_t}<0,\quad s_t=0,\qquad t=1,\ldots,F;$$

$$L_{\lambda_t}=I_0+\sum_{j=1}^{t}(x_j-s_j)>0,\quad \lambda_t=0,$$

$$or\quad L_{\lambda_t}=0,\quad \lambda_t\geq 0,\qquad t=1,\ldots,F-1;$$

$$L_\mu=I_0+\sum_{t=1}^{F}(x_t-s_t)=0.$$

These conditions are sufficient under the assumed concavity of all the R_t, and convexity of C.

We first find a solution which ignores the nonnegativity constraints on sales, output, and inventory, but which preserves the condition $I_F=0$. That solution will have the structure

$$(4)\quad C'(x_t)=R_t'(s_t),\qquad t=1,\ldots,F$$

$$(5)\quad C'(x_t)=\theta^{-1}[C'(x_{t-1})+h]$$
$$=\theta^{1-t}\left[C'(x_1)+h\sum_{j=1}^{t-1}\theta^{-j}\right]$$

$$(6)\qquad I_0=\sum_{t=1}^{F}(s_t-x_t).$$

In every period, marginal sales revenue equals marginal production cost. Production is arranged through time in such a way that the marginal unit produced in this period costs as much as the marginal unit produced last period and carried into this period. And the final inventory constraint holds.

By the monotonicity of C', we have from (3) that

$$(7)\quad x_t=(C')^{-1}\left\{\theta^{1-t}\left[C'(x_1)+h\sum_{j=1}^{t-1}\theta^{-j}\right]\right\}$$

A solution, ignoring the nonnegativity constraints, then can be obtained by choosing a value of x_1, which immediately by (7) implies values of $x_2,\dots,x_F$, and by (4), values of $s_1,\dots,s_F$. The search is for an x_1 which satisfies (4), (7), and (6). The search is straightforward because C' and R' are monotonic. If, upon completion of this search, it is found that x_t and $s_t \geq 0$, $t=1,\dots,F$, and $I_t \geq 0$, $t=1,\dots,F-1$, then the optimization problem is solved. The optimal horizon T coincides with the forecast horizon F.

However, suppose that one or some combination of those constraint conditions is violated. We rule out $s_t < 0$ and $x_t < 0$ on these grounds: our solution would call for $s_t < 0$ for $t < L$ only on occasions when demand takes such a large upward jump in period L that it swamps all earlier demands in its profit implication. Setting sales negative in periods up to L is a mechanical response whereby inventory entering that period of extraordinary demand can be made adequate (marginal production costs must be sharply rising to induce those negative sales in earlier periods). We rule out this possibility as untypical. Negative outputs $(x_t < 0)$ can occur if beginning inventory I_0 is of the same or greater magnitude than the total revenue-maximizing sales sequence. Given that the zero terminal inventory for this decision sequence is the beginning inventory of the next decision sequence, that possibility cannot be recurrent and so we ignore it.

We are left with the possibility of negative inventories within the planning period: $I < 0$, $\lambda > 0$ in one or more periods between 1 and F.

Suppose that, associated with the constraint-ignoring solution satisfying (4)–(7), negative values of inventory are observed in (say) period q, then later in period r. The response is to impose condition (6), summing to q, not F; this is done by increasing x_1 (and hence by (7), $x_2,\dots,x_q$), and reducing $s_1,\dots,s_q$, preserving the conditions (4). At period $q+1$, a new output and sales sequence is begun with beginning inventory $I_q = 0$; that sequence terminates in period F with zero inventory. Again, in some intermediate periods between q and F, negative inventories might be observed; they would certainly occur in r if anywhere. The first of those negative inventories is eliminated, just as $I_q < 0$ was corrected. The sequence up to period q, resulting in $I_q = 0$, will be part of an optimal and feasible program over $1,\dots,F$, and q will demarcate the optimal horizon T, only if the condition

$$C'(x_{q+1}) \leq \theta^{-1}\left[C'(x_q)+h\right] \tag{8}$$

holds after a feasible program has been developed over the entire planning interval. If (8) does not hold, it will pay to increase x_1, maintain the first-order conditions (4) and (7), and pass through period q in such a way that $I_q > 0$. At some later period T before period F, a zero-inventory condition will occur, and the condition

$$C'(x_{T+1}) \leq \theta^{-1}\left[C'(x_T)+h\right] \tag{8'}$$

will be satisfied for that period in the optimal program. A reduction in output in some periods beginning with $T+1$ will be necessary in order to satisfy the terminal inventory condition (6); that output reduction assures that (8′) will hold. Once the condition (8′) is established at $T=q$ or later, an optimal horizon is located at T.

Properties of optimal behavior can be studied by reference to that truncated program. Hereinafter we will assume that the optimal horizon T has been located, and we will focus on the firm's optimal decisions in its program over the T-period horizon. We will thereby be spared further concern over satisfying the constraints (2).

The existence of an optimal horizon is not a useful fact to one who is concerned with solving real world programming problems; as we have seen, the optimization routine must be carried out as far into the future as available information permits before the presence of the optimal horizon can be located with complete assurance. However, knowledge that the horizon exists, and of the conditions that must hold up to and at the horizon, is useful in exercises like this one, which explore the properties of optimal solution sequences.

Because of the first-order conditions and constraints that must hold in and between all periods up to an optimal horizon, a typical intrahorizon pattern will involve a sequence of low-demand periods during which inventory is accumulated, followed by a sequence of a higher-demand periods during which inventory is sold. The horizon will occur only between two periods in which there is a reduction in demand, or a slowdown in the rate of increase of demand. If the firm faces a pattern in which the demand rate peaks in the first period, then a horizon will occur at the end of the first period (unless initial inventory is high), and the remaining periods will be a separate planning problem.

Although full certainty, the model we use here resembles Blinder's certainty-equivalent formulation. In neither case does the variance or any higher-order moment of future demand play any role in determining optimal behavior. His objective function is quadratic in revenue and sales: with notation changed to resemble ours in the following, it is

$$\operatorname{Max} E\pi = E\sum_{t=1}^{\infty}\theta^{t-1}\big[s_t(\alpha-\beta s_t+\varepsilon)-ax_t$$

$$-b(x_t-\tilde{x})^2-h_1I_t-(h/2)I_t^2\big],$$

where ε is an identically independently distributed or $AR(1)$ process, and α, β, a, b, h_1, and h are demand and cost parameters. Because of the quadratic structure of the objective, only the mean value of future ε's enter the calculation; however, the program must be recomputed every period to take account of the immediate past period's realization of ε. Thus Blinder's model, which stands in precisely the same relation to Holt et al. as ours does to the earlier Modigliani-Hohn, is addressed to the problems that arise when the principal element of demand fluctuation is its error component.

II. Price Stickiness

We are now ready for the central result of this paper.

THEOREM 1: *In an output smoothing firm, price stickiness can be observed when demand is foreseen to decrease. It will never be observed when demand is foreseen to increase.*

PROOF:

Let the firm operate according to an optimal plan over the optimal horizon interval $[1,T]$. Let t and $t+1$ be adjacent periods within the interval, and let $D_t = D_{t+1}$, that is, the demand curve is the same in periods t and $t+1$. From the first-order conditions (4) and (5), the realized values of marginal revenue must satisfy

$$R'(s_{t+1}^*) = [R'(s_t^*)+h]/\theta, \tag{9}$$

where the asterisks denote optimal values. Because the R' functions are the same, $\theta<1$, and $R''<0$, it must be that $s_{t+1}^* < s_t^*$; and hence price must increase between periods: $p_{t+1}^* > p_t^*$. That tendency for prices to increase can be manifested as sticky pricing behavior. Suppose that demand shifts downward from period t to period $t+1$, so that the marginal revenue curve in $t+1$ lies below the marginal revenue curve in t. But even with that shift, we see from (4) and (5) that necessarily

$$R'_{t+1}(s_{t+1}^*) = [R'_t(s_t^*)+h]/\theta,$$

which in turn necessarily implies that $s_{t+1}^* < s_t^*$, and hence possibly that $p_{t+1}^* > p_t^*$. This possibility—a downward demand shift accompanied by higher price—has no symmetric counterpart. Under no circumstances will we find $p_{t+1}^* < p_t^*$ following from an upward shift in demand.

This result illustrates that within an optimal planning horizon, there is a tendency for price to increase through time. As we have already seen from the first-order conditions, the optimal plan within a horizon calls for the accumulation of inventory during earlier periods when demand is low, and the drawing down of inventory later when demand is high. Small "local" downward shifts in demand during the time of inventory accumulation may in the optimal plan evoke no reduction in price, only a greater rate of accumulation of inventory.

III. The Magnitude of Stickiness in a Linear-Quadratic Firm

To quantify and further clarify the implications of this tendency for price to drift upward within a horizon interval, we proceed with an example involving a linear demand curve and a quadratic and linear cost structure.

At the beginning of period 1, a production and sales schedule is implemented up to the firm's optimal horizon at the end of period T. Production and/or sales rates are changed at the beginning of each period, for example, Monday morning every week. Period t's price is given by a demand structure of the form $p_t = \alpha_t - \beta s_t$ where β is the temporally constant slope of the demand curve, and shifts in demand come about through changes in the intercept α.

Total production cost in period t is $ax_t + b(x_t - \tilde{x})^2$; $x_t \geq 0$. This specification is a general quadratic which explicitly embodies the minimum point of average production cost, $\tilde{x}$. The coefficients a, b, and $\tilde{x}$ do not change through time: the analysis is short run. There is no cost of changing the rate of production from period to period.

With an initial inventory holding of $I_0 \geq 0$, the period t ending inventory is $I_t = I_0 + \Sigma_{j=1}^{t}(x_j - s_j)$ and inventory-holding cost is $h \cdot I_t$.

To reduce clutter, let the discount factor θ equal one.[5] The objective is

$$(1') \quad \operatorname{Max} \pi = \sum_{t=1}^{T} \left\{ (\alpha_t - \beta s_t)s_t - ax_t - b(x_t - \tilde{x})^2 - h \cdot I_t \right\}$$

and $s_t \geq 0$, $x_t \geq 0$ $(t = 1,\dots,T)$, $I_t \geq 0$ $(t = 1,\dots,T-1)$, and $I_T = 0$ are satisfied because T is the optimal horizon.

The first-order conditions are

$$(10) \quad \pi_{x_t} = -a - 2b(x_t - \tilde{x}) - h(T - t + 1) = 0$$

$$(11) \quad \pi_{s_t} = \alpha_t - 2\beta s_t + h(T - t + 1) = 0$$

for $t = 1,\dots,T$; and we also have that

$$(12) \quad I_0 + \sum_{t=1}^{T} (x_t - s_t) = 0.$$

The relations (10), (11), and (12) imply that

$$(13) \quad x_1 = (\bar{\alpha} + 2b\tilde{x} - a)/2(b+\beta) - \beta I_0/T(b+\beta) - (T-1)h/4b$$

$$(14) \quad \Delta x_t = x_{t+1} - x_t = h/2b$$

$$(15) \quad s_1 = (2b\tilde{x} - a)/2(b+\beta) + \alpha_1/2\beta - \bar{\alpha}b/2\beta(b+\beta) + bI_0/T(b+\beta) + (T-1)h/4\beta$$

$$(16) \quad \Delta s_t = (\Delta\alpha_t - h)/2\beta$$

$$(17) \quad p_1 = \alpha_1/2 - \beta(2b\tilde{x} - a)/2(b+\beta) + \bar{\alpha}b/2(b+\beta) - \beta bI_0/T(b+\beta) - (T-1)h/4$$

$$(18) \quad \Delta p_t = (\Delta\alpha_t + h)/2,$$

where $\bar{\alpha} = T^{-1} \sum_{t=1}^{T} \alpha_t$.

The expressions (13)–(16) generate sequences x and s which satisfy the first-order conditions (10) and (11) and the constraint (12).

To better isolate the effect of inventory holding insofar as price movements are concerned, we compare the price histories of two firms with identical intertemporal linear demand structures and identical quadratic costs of production. One firm holds inventory, with the consequences exhibited in (13)–(18), while the other neither holds inventory nor sustains any inventory-associated out-of-pocket costs.

In a linear-quadratic firm which cannot hold inventory, the objective function is (1'), with $x_t = s_t$ and the cost-of-inventory term

[5] From (9) we see that inclusion of discounting would only increase the magnitude of price stickiness observed.

deleted. The optimal price change rule is

(19) $\Delta p_t^{\omega} = (2b+\beta)\Delta\alpha_t/2(b+\beta).$

(This follows immediately from the analogue of equation (10).) Comparing the interperiod price changes induced by interperiod movements in α in the two cases, we see from (19) that, without inventory, the price change always takes the same sign as the change in α, and the magnitude of Δp_t^{ω} is always greater than the magnitude of $\Delta\alpha_t/2$. With inventory holding as a part of the optimal program:

(i) a shift upward in demand always yields a price increase larger than $\Delta\alpha_t/2$;

(ii) a small downward shift in demand (with $|\Delta\alpha_t| < h$) yields a small price increase;

(iii) a larger downward shift in demand (with $|\Delta\alpha_t| > h$) yields a price reduction, but one smaller in magnitude than $|\Delta\alpha_t|/2$.

Thus, prices respond more sluggishly to a reduction in demand than they do to an increase in demand, and price may exhibit sticky behavior when the demand reduction is small.

IV. Replication and Extension of Blinder's Major Results

Our model straightforwardly yields Blinder's important conclusions. This dispels the idea that linearity of the inventory-holding cost function is limiting, or leads to unusual patterns of optimizing behavior, as he finds; more positively, it offers reassuring evidence regarding the robustness of his results.

Notationally, we have already used $\Delta\alpha_t$, Δs_t, etc. to denote shifts in exogenous or controlled variables that occur between periods. We now introduce $\delta\alpha_t$ as a change from the forecast value of α *within* period t, this change being foreseen at time $t-1$ (the beginning of period t) or earlier, but after time 0; or perhaps not being foreseen at all. This type of change seems to conform reasonably to the notion of a "shock".

If such intraperiod demand changes are foreseen before they occur, the firm in effect is faced by a new optimization problem based on a new forecast of demand coefficients: whereas the old forecast sequence was $\{\alpha_t\}$, the new forecast sequence is $\{\alpha_t + \delta\alpha_t\}$. If the changes occur subsequent to time J and they are foreseen at time J, then the new problem is

$$\operatorname*{Max}_{\substack{x\geq 0\\ s\geq 0}} \sum_{t=J+1}^{T} \left\{ [(\alpha_t+\delta\alpha_t)-\beta s_t]s_t - ax_t - b(x_t-\tilde{x})^2 - h\cdot\left[I_J + \sum_{j=J+1}^{t}(x_j-s_j)\right]\right\}$$

subject to $I_T = 0$. We assume for now that the vector of shocks $\delta\alpha_t$ does not displace the original optimal horizon T.[6]

Alternatively, if an unforeseen shock occurs in period J, and no shocks are foreseen subsequent to J, the objective is

$$\operatorname*{Max}_{\substack{x\geq 0\\ s\geq 0}} \sum_{t=J+1}^{T} \left\{ (\alpha_t-\beta s_t)s_t - ax_t - b(x_t-\tilde{x})^2 - h\cdot\left[I_J + \delta I_J - \sum_{j=J+1}^{t}(x_j-s_j)\right]\right\}$$

subject to $I_T = 0$.

From (13)–(18) we discover that, with foreseen shifts in demand during periods $J = 1,\dots,T$, the responses in those periods are

(20) $\delta x_t = \delta\bar{\alpha}/2(b+\beta)$

(21) $\delta s_t = \delta\alpha_t/2\beta - b\delta\bar{\alpha}/2\beta(b+\beta)$

$\delta p_t = \delta\alpha_t/2 + b\delta\bar{\alpha}/2(b+\beta)$

where

(22) $\delta\bar{\alpha} = \sum_{t=J+1}^{T}\delta\alpha_t/(T-J).$

Note that output does not change when $\delta\bar{\alpha} = 0$. If subsequent demand shifts sum to zero, that is, if demand is rearranged through

[6]A displacement could occur if the $\delta\alpha_t$ induced larger sales, and caused inventory to go negative before period T, for example.

time but unchanged in total quantity, then there is no adjustment of the output schedule. Blinder asserts that demand shocks will not affect output if inventory-holding cost is linear; only shock patterns which satisfy the condition $\delta\bar{\alpha}=0$ will conform to his finding.

Otherwise, when there is an increase in total demand from the foreseen pattern, output increases in every period, sales increase in periods when $\delta\alpha>0$ and diminish in periods when $\delta\alpha<0$, and increase overall; and price has added impetus to rise: only when $\delta\alpha$ is negative and sufficiently large in absolute magnitude is price lower than in the original schedule.

When a demand shock is not foreseen its effect is transmitted through I_J. We see from (13)–(18) that for $t=J+1,\dots,T$,

$$\delta x_t=-\beta\delta I_J/[(b+\beta)(T-J)], \tag{23}$$

$$\delta s_t=b\delta I_J/[(b+\beta)(T-J)], \tag{24}$$

$$\delta p_t=-b\beta\delta I_J/[(b+\beta)(T-J)]. \tag{25}$$

When continuing inventory is smaller than originally planned ($\delta I_J<0$), output and price are increased and sales reduced in every period, compared to the original optimal plan. This does *not* mean that a positive value δI_J always yields an actual price reduction from period $J-1$ to period J; it only means that the revised p_J is smaller than was planned at the beginning of the optimal horizon.

If the optimal horizon is lengthened or shortened by the pattern of demand shocks, the effects can be seen directly in (20)–(25), by substituting T^* (the new value of the optimal horizon) for T.

For the instances analyzed here, with T unchanged, we have seen the following.

(a) A positive-valued total shift in demand during period J causes an increase in price, output, sales, and a reduction in inventory investment x_J-s_J. This result is identical to Blinder's Proposition 1.

(b) Per his Proposition 3, the more persistent the shock (*ceteris paribus*, the larger the increase in $\delta\bar{\alpha}$), the greater the responsiveness of output and price and the less the responsiveness of sales and inventory investment.

Blinder's Proposition 2 relates the responsiveness of sales and price to changes in the storability of output, measured by the curvature of his quadratic inventory holding cost function. We find it more natural to interpret the *level* of marginal holding costs as a measure of the storability of output, but neither model yields easily interpretable comparative dynamic results with respect to that storability measure.

V. Interpretive Remarks and Caveats

In Section II it was seen that the firm's response to a shock-free, perfectly foreseen pattern of demand fluctuation can indeed generate price stickiness, using Definition 1. And although demand shocks (as defined in Section IV) will displace the optimal price sequence that was originally planned over the horizon, the displacement caused by a demand-reducing shock may not be sufficient to bring about an observed price reduction in subsequent periods.

How reasonable is it to suppose that inventory management as in Section II offers a primary explanation of price stickiness? The conditions (*i*)–(*iii*) of Section III suggest that to find out, we compare the magnitudes of interperiod changes in α, the intercept of the linear demand curve, and h, the marginal and average cost of holding a unit of finished goods in stock for one period.

In the short run, the principal component of storage cost h is the interest charge on inventory investment. Other elements include spoilage or pilferage (to the extent not covered by insurance), obsolescence (as with style or seasonal goods), handling costs (which probably rise less than linearly with quantity stored), and taxes and insurance (either of which may be more in the nature of overhead costs than direct costs). A firm may also elect to hire storage space rather than invest in its ownership; symmetrically, a firm that owns storage space has the option of letting it. Depending on the nature of the space lease contract, actual or imputed charges for space could be another component of the direct cost of holding finished

goods. The capital costs of storage space are borne by the firm, and the more nearly perfect the rental market for space, the more nearly it is true that the capital cost varies directly with the level of the firm's own inventory.

We assume that this market for storage space is imperfect, so rental contracts call for fixed amounts of space over long periods of time. Storage space costs then do not enter into the holding coefficient h. The magnitude of h at current interest rates might then be approximately 20 to 25 percent of average production cost, calculated on an annual basis. If the scheduling interval is a month, h is on the order of 1.6 to 2.1 percent of unit production cost; if a week, .38 to .48 percent.

Suppose that in our monopoly firm the intercept α is on the average twice as large as market price, and that market price is about twice the average production cost. The intercept is then sixteen to twenty times the annual holding cost. Then any weekly demand reduction (shift in α) smaller than .1–.12 percent, or any monthly reduction smaller than .41–.52 percent, will cause no price reduction whatever if the discount factor θ is close to one. These are not large reductions in demand; if they persist, they cumulate to only 5 to 6.25 percent over the course of a year. These magnitudes, which cannot plausibly be vastly increased, make it appear to be well worth looking beyond inventory-holding behavior for additional explanations of price stickiness in the monopoly firm.

These are the immediate implications of the comparison of $\Delta\alpha$ to h, all *ceteris paribus*:

1) Prices will respond more when demand increases than when demand falls.

2) The flatter the demand curve and the smaller the price-cost markup, the more sluggish and sticky prices are.[7]

3) The higher the interest rate, the stickier prices will be when demand drops, and the less sluggish prices will be when demand increases.

4) Assuming maximization of expected returns, the probability of obsolescence, spoilage, or pilferage of stored goods affects price response just as the interest rate does.

5) A rise in storage space rents (for example, in response to increasing site values), a decrease in the "thinness" of the market for storage space, or an increase in the sophistication of the contracts to let such space, all affect price stickiness/sluggishness in the same way as a rise in the interest rate.

6) The prices of goods that are particularly hard to handle or expensive to store because of weight or bulk per dollar of value will respond as though to a high interest rate.

Four caveats seem to be worth recording. First, our analysis, like Blinder's, focuses on the pricing policies of manufacturers. A model of behavior at the retail level involves an entirely different analytic structure, largely because the retailer is concerned with costs of holding and procurement, not with costs of holding and production. If we include a retail trade sector, or any "multistage" structure of inventory maintenance (Scarf, Dorothy Guilford, and Maynard Shelley, 1963), we should note that the likely effect of inventory maintenance at the retail level will be to amplify short-term fluctuations in consumer demand before transmitting them to the manufacturer (Orr, 1963). Longer-term swings should not be affected appreciably by retailer's holdings. But the increase in short-term fluctuation caused by decisions at the retail level weakens inventory cost as a primary explanation of stickiness in manufacturers' pricing behavior, because it makes the uncertainty buffering motive more important and the smoothing motive less important.

That leads to our second caveat: the results we obtain here, which include actual downward stickiness of prices, stem not only from our linear holding cost assumption but also from our full-certainty assumption. It is not easy to move into a stochastic demand setting with linear holding costs, because the nonnegativity constraints on inventory rule

[7]This condition is a peculiar one, and indicates that we should be mistrustful of results which refer to "competition" when the demand curve in a monopoly model approaches the horizontal. A true competitor has no control over price, while a monopolist with highly price-sensitive demand will exhibit considerable stickiness.

out certainty equivalence.[8] Without those constraints linear holding costs lead to infinitely large order backlogs, and without certainty equivalence, stochastic models yield explicit solutions only grudgingly.

By way of contrast to our full-certainty model, in a more explicitly stochastic setting the decision sequence would be reviewed once per period and revised to compensate for the realized stochastic demand component in the period just ended, which affects initial inventory holdings in the current period; and also to compensate for revisions in demand forecasts. If the revised plan always looks a constant number of periods into the future, or if in each period the planning problem is reformulated with an infinite horizon, the firm will always be in the first period of a plan of unchanging duration. In that case the effect of inventory on price that was a central element in the proof of Section II will be swamped, so that the downward price stickiness noted there no longer can be expected to occur. However, if the firm plans toward an unchanging horizon, as might happen with a strong seasonal peak, the transient effects could be small relative to the upward price drift which accompanies movement toward the horizon, and stickiness might be observed.

Third, the focus on finished goods inventory, with the exclusion of materials and work in process, is a convenience in modeling, but it may occasionally give rise to deceptive conclusions. The more tightly that finished goods inventory is controlled, that is, the less variation there is in inventory level or usage, the more we expect to see the output rate fluctuate. Wider fluctuations in the output rate may call for higher average holdings of materials or work in process. It is not proper to treat inventory at all stages as a single capital good, because the inventory conservation identity $I_t = I_{t-1} + x_t - s_t$ cannot hold in physical unit terms if I_t contains stock at all levels of completion; nor, even after allowance for purchases and value-added at different stages, is there any admissible single-constraint representation of inventory conservation in value terms. To accommodate inventory at n stages of completion, it is necessary to include n constraint equations.

Fourth, we note that it has for the most part been macroeconomists who have incorporated price stickiness in their modeling exercises, and who have observed prices or wages to be sticky. Individual choice behavior, in this paper and elsewhere, has been invoked to explain stickiness; but even if a robust and convincing explanation of price stickiness can be found at the micro level, additional theorizing or empirical work will be necessary to carry the explanation forward to cover stickiness in aggregate data. For even though the pricing behavior of individual firms may exhibit stickiness—long periods of no change, punctuated by occasional significant changes—any aggregate index of their prices will change more or less continuously, unless the price changes of the individual firms are somehow temporally coordinated (as by price leadership, or by common sensitivity to the price of a primary input resource). The larger the number of firms and the less coordinated their actions, the less aggregate stickiness will be evident in such an index, regardless of how sticky individual firms' prices may appear to be.

The model we have used here is probably the most supportive of any monopoly model that can be found for the view that price stickiness is a product of inventory management decisions. We have given that view its best chance and at this stage it appears to be unsupportable. Further modeling activity might examine the consequences of oligopolistic interaction.

[8]See Holt et. al., ch. 6.

REFERENCES

Arrow, Kenneth J., Karlin, Samuel and Scarf, Herbert, *Studies in the Mathematical Theory of Inventory and Production*, Stanford: Stanford University Press, 1958.

Blinder, Alan S., "Inventories and Sticky Prices: More on the Microfoundations of Macroeconomics," *American Economic Review*, June 1982, *72*, 334–48.

Charnes, Abraham, Drèze, Jacques and Miller, Merton H., "Decision and Horizon Rules for Stochastic Planning Problems: A Linear Example," *Econometrica*, April 1966, *34*, 307–30.

Hadley, George and Whitin, Thomson M., *Analysis of Inventory Systems*, Englewood Cliffs: Prentice-Hall, 1963.

Holt, Charles C. et al., *Planning Production, Inventories and Work Force*, Englewood Cliffs: Prentice-Hall, 1958.

Lee, Dwight R., "Dynamic Profit Maximization Under Conditions of Monopoly," unpublished doctoral dissertation, University of California-San Diego, 1972.

______ **and Orr, Daniel,** "Further Results on Planning Horizons in the Production Smoothing Problem," *Management Science*, January 1977, *23*, 490–98.

Means, Gardner, *The Roots of Inflation: The International Crisis*, New York: Burt Franklin & Co., 1975.

Modigliani, Franco and Hohn, Franz M., "Production Planning Over Time and the Nature of the Expectation and Planning Horizon," *Econometrica*, January 1955, *23*, 46–66.

Orr, Daniel, "A Stochastic Income Model with Optimal Inventory Rules," *Review of Economic Studies*, April 1963, *30*, 84–92.

Scarf, Herbert E., Gilford, Dorothy M. and Shelley, Maynard W., *Multistage Inventory Models and Techniques*, Stanford: Stanford University Press, 1963.

Schutte, David P., "Inventories and Sticky Prices: Note," *American Economic Review*, September 1983, *73*, 815–16.

Shaw, Edwin S., "Elements of a Theory of Inventory," *Journal of Political Economy*, August 1940, *48*, 465–85.

Costly Employment Contract Renegotiation and the Labor Mobility of Young Men

By JOHN J. ANTEL*

The distinction of quits and permanent layoffs remains a contentious issue in the labor economics literature. Basically, there appear two contrasting views of such turnover. While both interpretations of turnover are derived from job-matching models that assume wealth-maximizing behavior of workers and firms, each involves different assumptions about the feasibility of individual labor contract negotiations following hire. On the one hand, application of the Coase theorem to employment contracts implies that turnover will occur only when the value of the worker's product in some alternative employment exceeds current job productivity. Since the actual division of the job match rent is assumed costlessly renegotiable, only the presence of a truly more productive job assignment would prevent the renegotiation of the current rent-sharing contract to the mutual benefit of both worker and firm. The likelihood of a job change is thus directly related to the level of total match productivity irregardless of whether the worker, in response to some later survey question, characterizes the separation as a quit or permanent layoff. According to this formulation of worker and firm attachments, quits and permanent layoffs are, at least in any sense relevant to actual resource allocation, indistinguishable. An implication of this model is of course that mobility enhances productivity (see Gary Becker, Elisabeth Landes, and Robert Michael, 1977).

Counter to this view stand models of job matching that suggest that the presence of transactions costs in the form of costly contract renegotiations preclude any easy application of the Coase theorem to employment contracts. Most forcefully advocated by Masanori Hashimoto (1981), these models of job matching imply that quits and permanent layoffs are different. The difference results from the fact that, with costly contract renegotiation, mobility is to a large extent determined by the actual rent-sharing agreement formulated at the beginning of employment. According to this view then, the total value of the match does not exclusively determine job change, and thus not all mobility is productivity augmenting.

This paper offers a model of job matching with costly post-hire negotiations in some aspects similar to that of Hashimoto, but with particular relevance to the population of young workers only beginning their labor force participation. While Hashimoto's discussion concentrates on longer-term firm-specific training decisions, the current model focuses on the period immediately following hire when worker productivity is to a large extent governed by endowed capabilities rather than determined by learning on the job. The theoretical focus on inexperienced workers as well as the corresponding youthful composition of the empirical sample are motivated to control for unobserved factors such as firm-specific human capital stocks as much as possible.

In further contrast to Hashimoto's emphasis, the model yields empirical implications concerning the role of wages in the determination of mobility that contrast to the implications of an otherwise similar zero negotiations cost model of job matching. Derivation of the empirical implications concerning the relationship of wages to quits and layoffs represents the primary contribution of this paper. Finally, empirical results are presented that, at least for a sample of very young men over a short period of time,

*Department of Economics, University of Houston, Houston, TX 77004. I thank Finis Welch, Dan Mitchell, and the referee for comments on an earlier version of this paper.

tend to support the costly negotiation view of employment contracts and turnover.

I. Theoretical Model

This section outlines a model of job matching with costly contract renegotiation that generates quits and permanent layoffs as a result of both match rent-sharing agreements initiated at the beginning of employment and as the by-product of worker and firm learning behavior after initial hire.[1] This model is then contrasted to a similar model without costly renegotiation of employment contracts. Both models apply to a population of inexperienced young men. Concentration on this age group permits us as much as possible to abstract from the effects of accumulated specific training, skill obsolescence, and previous job experiences on turnover decisions.

A central assumption is that individual worker productivity varies with firm assignment, and that individual workers differ with respect to their comparative advantage in employment. It is further assumed that job-specific worker productivity is imperfectly known either to the worker or to the firm prior to some trial period of work. In summary, individual productivity is match-specific and an "experience good." These assumptions are consonant with the youthfulness and inexperience of the population under study: young workers are relatively unknown and heterogeneous productive agents.

The model describes an employment contract that specifies at the time of hire how total match rents are shared between the worker and the firm. The primary behavioral assumption of the model is that workers and firms each try to maximize their own share of the rent stream derived from the match. At the time of hire, workers and firms are in agreement concerning both expectations of worker productivity and the likely value of alternative job offers. Thus, employment contracts based on these preliminary expectations are feasible. The rent-sharing contract is specified at the time of hire, but for reasons outlined below, renegotiation is limited for some time thereafter. Following hire, workers continue to receive other wage offers and firms become apprised of actual worker productivity on the current job. Quits result from the worker's receipt of a superior wage offer, while permanent layoffs result from the revelation to the firm of less than expected worker productivity.

In contrast to the assumptions of neoclassical theory, the "experience good" and match-specific nature of worker productivity imply that each worker's wage-offer distribution imperfectly approximates the distribution of his product over various jobs. This wage and productivity discrepancy follows for two reasons. First, since productivity is imprecisely understood prior to hire, even in the presence of competitive forces tending to equate wages and expected productivity, some initial-offer wages will include productivity estimation errors. Second, since productivity is assumed match-specific, in some cases match rent may be captured by the firm, thus implying wages below actual productivity.

Tractability mandated some model simplifications. First, workers and firms are assumed risk neutral. Second, workers are assumed unable to discriminate among firms on the basis of management ability to guess productivity. In fact, I sidestep completely any theoretical discussion of compensating wage differentials that may result from differences in layoff probabilities or other characteristics among firms.

Third, although workers search for better jobs and firms assess worker productivity during an initial trial period of employment, neither of these simultaneous learning processes is explicitly modeled. Rather, the results of job search comprising alternative wage offers for workers, and the result of

[1]The model was inspired in part by the work of Hashimoto (1979, 1981). Much of the notation and the arguments concerning asymmetrical information are derived from him. The models differ with regard to certain other assumptions and implications, however, since he assumes that the level of m reflects investments in firm-specific training. Hashimoto's model more closely resembles that of Donald Parsons (1972), although Parsons did not explicitly recognize the importance of transactions costs.

productivity monitoring comprising the firm's updated estimates of worker performance, are taken as first-period outcomes described by simple probability density functions. Quit and layoff decisions are then shown to follow from the actual realizations of these random variables. Such learning processes have been more specifically outlined in the work of others and will not be discussed here (see Steven Lippman and John McCall, 1976a, b; and Boyan Javanovic, 1979).

Finally, following Hashimoto, I maintain that the value of alternative job offers is known only to the worker as a result of job search, and further assume that the actual productivity of the match is revealed exclusively to the firm as a by-product of its normal production monitoring activities. This double asymmetry of information raises the possibility of various forms of opportunistic behavior, a potentiality which precludes post-hire renegotiation of wage contracts and deters the institution of various other incentive altering contingent payment schemes.[2]

The model comprises an initial trial period of employment after which workers either forever become tenured on the current job or switch to some alternative firm. At the beginning of the trial period, the worker and the firm are assumed to maintain consonant expectations concerning worker productivity on the current job and worker accessibility to alternative employments. These expectations are similar enough to allow the negotiation of an employment contract that stipulates how the expected match rent will be shared by the employer and employee throughout the lifetime of the job match. Workers receive a wage payment,

$$w = H + \alpha m, \tag{1}$$

where H represents the market value of general human capital, and $0 \leq \alpha \leq 1$ is the fraction of the expected match rent m maintained by the worker. The firm receives the residual share of the expected rent,

$$(1-\alpha)m. \tag{2}$$

After the first period, receipt of these rent shares is conditional upon the effective reaffirmation of the contract by both parties —the worker does not quit and the firm does not permanently layoff the worker.

In the negotiations prior to hire, worker and firm attempt to fix α in order to maximize the present value of their respective rent-share streams. Workers attempt to maximize the objective function,

$$\phi_w = H + \alpha m + 1/r\{Q\cdot E(\varepsilon|\varepsilon > \alpha m) + (1-Q)[L\cdot E(\varepsilon|\varepsilon \leq \alpha m) + (1-L)\alpha m]\}, \tag{3}$$

while the firm's objective function is given by

$$\phi_f = (1-\alpha)m + 1/r\{(1-Q)(1-L) \times [(1-\alpha)m + E(\eta|\eta > -(1-\alpha)m)]\}. \tag{4}$$

The expressions Q and L represent quit and layoff probabilities:

$$Q = \int_{\alpha m}^{\infty} g(\varepsilon)\,d\varepsilon, \tag{5}$$

$$L = \int_{-\infty}^{-(1-\alpha)m} f(\eta)\,d\eta. \tag{6}$$

Each objective function represents the sum of the trial period payoff plus the discounted at rate r present value of the future wage stream weighted by the respective separation probabilities.

[2] For example, the practices of internalizing externalities accompanying mobility by the payment of "damages" or matching outside offers, both so skillfully discussed by Dale Mortensen (1978) in the context of a matching model with known total productivity, are not feasible under the present informational assumptions. Asymmetrical information implies that firms cannot be sure of the veracity of outside offers, nor is any amount of damage resulting from quits or terminations public information. Other cost considerations could limit renegotiations. Firms may prefer to negotiate only at specified off-peak production times. Worker and firm learning out of phase might imply some limit to renegotiation if, for example, the worker received a superior offer before the firm could accurately assess productivity.

Expressions (5) and (6) indicate that, conditional on α and m, quits and layoffs are determined by the values of the random variables η and ε. The nonnegative random variable ε indicates the per period value of the worker's postsearch best alternative job rent share, while the random variable r represents the deviation of the true productivity on the current job from its initial expectation m. The actual value of ε is revealed exclusively to the worker after the first job period. The value of ε derives from the worker's job search activity. The value of η is revealed exclusively to the firm following the observation of worker performance during the trial period. The firm is an unbiased guessor of productivity so that $E(\eta)=0$. Functions $g(\varepsilon)$ and $f(\eta)$ describe the probability densities of these random variables that are assumed independent of each other and also independent of H, α, and m. The densities thus describe the outcomes of two simultaneous, but separate, firm and worker-learning processes occurring during the initial trial period of employment. The model implies that receipt of a large positive ε by the worker elicits a quit, while discovery of a large negative η by the firm evokes a permanent layoff.

Although I abstract from any specific discussion of the bargaining process determining α in pre-hire negotiations, some bounds may be placed on the Pareto optimal values of α within the unit interval. For the purpose of deriving these bounds, worker and firm objective function first derivatives with respect to α are given below with the signs of each component given in the underlying parentheses.

$$(7)\quad \frac{\partial \phi_w}{\partial \alpha}=\underset{(+)}{m}+\frac{m}{r}\Big\{f(-(1-\alpha)m)\times\underset{(-)}{(1-Q)\,(E(\varepsilon|\varepsilon\le \alpha m)-\alpha m)}+\underset{(+)}{(1-Q)\,(1-L)}\Big\},$$

$$(8)\quad \frac{\partial \phi_f}{\partial \alpha}=-\underset{(-)}{m}+\frac{m}{r}\Big\{g(\alpha m)(1-L)\times\Big[\underset{(+)}{(1-\alpha)m+E(\eta|\eta>-(1-\alpha)m)}\Big]-\underset{(-)}{(1-L)(1-Q)}\Big\}.$$

Although the signs of the first and last components of these derivatives are consistent with the intuitive notion that workers prefer α larger while firms prefer a smaller α, the sign of the second component in each expression implies that workers and firms may be restrained in their choice of maximal α by the prospect of increasing the probability of some match dissolving action precipitated by the opposite party. For example, referring to the second term in (7), we see that when both the worker's expected net loss from a layoff, or $(1-Q)(E(\varepsilon|\varepsilon\le\alpha m)-\alpha m)$, and the sensitivity of L to changes in $-(1-\alpha)m$ given by $f(\eta)$ are large, then (7) may become negative for some larger values of α. In essence, the probability of layoff may become too high for large values of α. This implies that workers may optimally choose an α less than one. Similarly, (8) may become positive if the firm expects a large return on the match and the sensitivity of Q to changes in αm is large for some lesser values of α. This latter possibility implies that the firm may prefer a value of α greater than zero. An example of such interior maximums is given in Figure 1.

The model as it stands does not imply a unique value of the rent-sharing parameter. However, the Pareto optimal set of α's is given by that area bounded by the two α values coincident with the individual objective function maximums. The final value of α will reside in this Pareto interval since values of α outside this region imply that both parties could be made better off by raising or lowering the parameter. To summarize the discussion concerning the determination of α, we may only conclude that, in the absence of some ancillary assumptions, that the designated value of α will lie in this Pareto

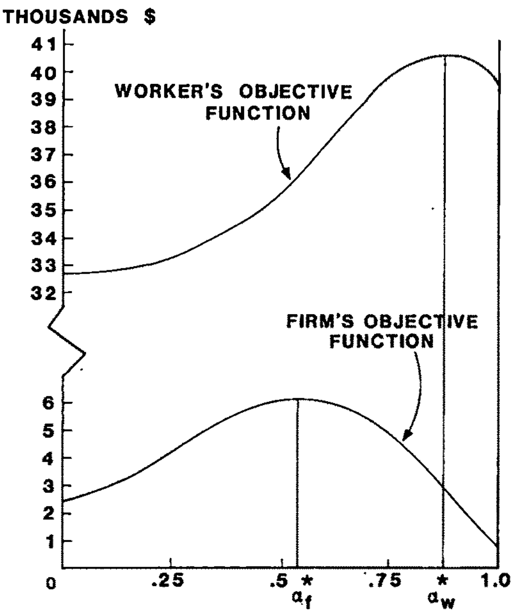


FIGURE 1

Note: ε and η are distributed normal with $\mu_\varepsilon = 1000$, $\mu_\eta = 0$, $\sigma_\varepsilon = 500$, $\sigma_\eta = 250$, $m = 2000$, $H = 2000$, and $r = .10$.

interval. In Figure 1, this set is the area between α_f^* and α_w^*.

Here again the present model contrasts with the work of Hashimoto in which workers are assumed to maximize joint wealth at the time of hire. According to Hashimoto's formulation, the initial employment contract is determined in two steps. First, the worker and firm maximize total joint wealth with respect to α and the second-period stock of firm-specific human capital acquired through training during the first employment interval. Second, the separate worker and firm wealth streams are determined by the allocation of first-period training costs among the contracting parties. In effect, this two-step aspect of Hashimoto's investment model separates decisions pertaining to the level of training or the distribution of second-period training payments from the determination of worker and firm expected values at the time of hire. The current model, however, does not involve training investment but rather takes productivity as given but imperfectly understood at the time of hire. Thus, joint wealth need not be maximized and, as in the classic bilateral monopoly case, some indeterminancy remains.[3]

Following agreement on some value of α, employment continues through the initial interval; and then, simultaneously at the end of the first job period, workers decide whether to quit or stay while firms decide whether to retain or layoff individual workers.

During the initial period, workers search over alternative employments and receive a maximal alternative rent-share offer described by the value of ε. If the new offer exceeds the current job value αm, a quit ensues. Since the worker alone invests resources in job search, I assume that only he is apprised of the true value of ε. The firm maintains its old expectations about the worker's alternative job prospects. The quit condition is thus

$$\varepsilon > \alpha m. \tag{9}$$

A layoff occurs at the end of the first period when the deviation of productivity, η, from its original expectation, m, is found negative enough to eliminate the firm's positive rent share. This condition is given by

$$\eta < -(1-\alpha)m. \tag{10}$$

Firms estimate the value of η as a by-product of management's general production monitoring and coordinating activities. This new productivity information is private to the firm, while workers maintain their original expectations about current job productivity given by m.

The privacy of the new productivity information to firms and the alternative job information to workers implies a double asymmetry of information that makes renegotiation of employment contracts prohibitively expensive. Incentives arise for firms to down-

[3]It is clear that the optimal social value of α also lies in this interval, although the proximity of the final α to this value cannot be determined. The model thus does not imply that contracts be *ex ante* efficient. In Hashimoto's model contracts are always, at least in expectation, socially efficient.

grade reports of worker productivity while workers will tend to exaggerate the value of alternative offers. These incentives for distortion and dissembling in post-hire negotiations force each party to preserve their initial expectations with respect to unobserved values of ε or η. Firms stick by their initial expectations with respect to the likely value of worker's alternatives and workers maintain their zero expectations with respect to η. Privacy of information precludes mutual readjustment of expectations which might serve as a basis for employment contract renegotiation. Renegotiation infeasibility implies that α, and thus wages, remain fixed. The fixity of α justifies the conditions determining job change suggested by equations (5) and (6), where the rent distribution parameter α and the expected productivity m in conjunction with the realizations of ε and η determine the probabilities of quit or permanent layoff.

It should be noted that the model does not assume either total firm ignorance of worker alternatives or complete worker unfamiliarity with productivity. Rather, it is assumed that although some public information exists which indicates current job productivity and alternative job opportunities, the value of this public information is fully exploited in pre-hire negotiations. And thus, both parties tend to maintain their partially informed initial expectations about the opposite party's more recent, but also private, information.

In contrast to the above turnover conditions implied by costly recontracting, the possibility of costless renegotiation suggests the Coase theorem type of result that both quits and permanent layoffs are determined at the end of the first period by the condition

$$(11) \qquad m+\eta<\varepsilon.$$

This inequality indicates that job change occurs only when a more productive match offer precludes a renegotiated value of α leaving some positive rent share for both parties. The conspicuous absence of α from expression (11) is rationalized by the fact that this parameter may be costlessly adjusted to reflect new information about ε or η at any time. The absence of α from (11) also implies that observed wages may only tenuously predict turnover. This "marriage model" of job matching represents the primary alternative to the model outlined at the beginning of this section. The contrasting empirical implications of the two models will be discussed subsequently. The contrasting welfare implications should be noted at this time, however.[4]

In contrast to the model with transactions costs, the model of job matching with costless renegotiation implies that job change always augments productive efficiency. According to this zero transactions cost model, rents exist, but the distribution of these rents among workers and firms doesn't affect turnover since the specific rent shares are continuously renegotiable. This leads to the Coase theorem type of result that, without transactions costs, job reassignments will be efficient since, otherwise, surrender of some rent share from one party to the other could maintain the match to the benefit of both.

Conversely, the model of job matching without post-hire negotiations does not imply that mobility always augments productive efficiency. In the absence of renegotiation, the actual rent distribution, and not total match productivity alone, determines quits and permanent layoffs. Quits occur when the worker, while searching on the job, receives a higher wage offer. However, since wage offers vary across different employers due to differences in rent-sharing agreements as well as arising from differences in expected productivity, a higher-paying job may not constitute a more productive employment.

Further, according to my model without renegotiation, permanent layoffs occur when productivity is found far enough below initial expectations to imply an actual loss in firm profits. That is, when the match rent

[4]For example, according to this model some "quits" may result from a decline in the current job offer anticipated by the worker following a downward estimation of worker productivity by the firm. Some "layoffs," on the other hand, may reflect the employer's anticipation of a future quit following the worker's receipt of a superior wage offer.

accruing to the firm is revealed to be negative following some trial period of work, a layoff ensues so that the firm may avoid a negative rent share. Again, this need not imply that the worker is more productive in some alternative employment, but only requires that the contracted wage in any subsequent job is low enough to leave some expected rent share for the new employer.

II. Empirical Implications

This section contrasts the empirical implications of the two competing models of job matching. The discussion centers on the relation of wages to job change. The section concludes with a brief discussion of other factors affecting mobility.

A. *Wages and Job Change*

In discussing the role of wages in the determination of mobility, it should be made clear at the outset that wages here represent survey wages observed at the beginning of some time interval during which each sampled individual is at risk to change jobs. Wages are not observed continuously at various jobs, but rather episodically at discrete times. The following discussion should be understood in this context.

Implications derived from the costly renegotiation model are discussed first. Referring back to expressions (1), (5), and (6), the following partial derivatives may be signed,

$$(12)\quad \frac{\partial W}{\partial m}>0,\quad \frac{\partial W}{\partial \alpha}>0,\quad \frac{\partial Q}{\partial \alpha}<0,$$
$$\frac{\partial Q}{\partial m}<0,\quad \frac{\partial L}{\partial \alpha}>0,\quad \frac{\partial L}{\partial m}<0.$$

According to this model, α and m both positively affect wages and both negatively relate to the quit probability. And thus, after controlling for the general human capital component of wages H, quits should be negatively related to wages. The value of the current job relative to other employments is increased either by a higher α or m, and an increase in either element of αm also tends to lower the likelihood of a quit.

Formally, the extent of the negative wage and quit relation is governed by the degree of correlation between the lower limit of the quit integral, αm, and the observed wage $H+\alpha m$, which given the independence of α, m, and after controlling for H, is obviously one: the specific component of the wage αm is a perfect proxy for the lower limit of the quit integral. The covariance component of this correlation is given

$$(13)\quad \operatorname{cov}(\alpha m,\alpha m)=\sigma_\alpha^2\sigma_m^2+\mu_\alpha^2\sigma_m^2+\mu_m^2\sigma_\alpha^2,$$

which describes the variance of αm.[5] Here the means and variances of the respective subscripted variables are given by μ and σ^2. Of special interest is the fact that only in the special case when the variance of α is zero are quits determined exclusively on the basis of productivity differences among jobs.

Permanent layoffs, however, are ambiguously related to the match specific component of wages αm, for although a higher m raises the firm's potential rent payment, a higher α also lowers the firm's actual rent share. Since α and m cannot be separately observed, the wage, or more precisely the match value component of wages, αm, ambiguously relates to the layoff probability. It may only be concluded that, after controlling for the H component of wages, an observed negative layoff and wage relation implies that m dominates the variance of αm, while a positive effect of wages on layoffs suggests the dominance of α. Thus, conditional upon the acceptance of the model, the observed relation of wages and layoffs is of some usefulness in determining the importance of α in the formulation of wage contracts. The model can offer no unambiguous predictions concerning the role of wages in the generation of layoffs, however.

[5]The derivation relies on the independence of α, m, and H. Starting with the formula, $\operatorname{Var}(\alpha m)=E((\alpha m)^2)-(E(\alpha m))^2$, and using the properties of expectation given independence, and substituting, $E(\alpha^2)=\sigma_\alpha^2+\mu_\alpha^2$, we may derive the formula for the variance of αm.

The effect of wages on permanent layoffs may be more formally exposited by deriving the covariance between the specific component of wages, αm, and the upper limit of the layoff integral, $-(1-\alpha)m$, which is[6]

$$(14)\quad \operatorname{cov}(-m+\alpha m,\alpha m) = \sigma_\alpha^2\sigma_m^2+\mu_\alpha^2\sigma_m^2+\mu_m^2\sigma_\alpha^2-\mu_\alpha\sigma_m^2.$$

In the extreme case of identical contract parameters across jobs, then α is constant between zero and one, and the expression becomes

$$(15)\quad (\mu_\alpha^2-\mu_\alpha)\sigma_m^2=(\alpha^2-\alpha)\sigma_m^2<0.$$

This expression is negative since α resides in the unit interval. In this extreme case, the variance of expected productivity m dominates the variance of αm, and layoffs, like quits, are a negative function of the observed wage after we control for H.

In the opposite extreme case where m is a nonzero constant, the covariance between the wage rent share and the expression determining the probability of a layoff is

$$(16)\quad \mu_m^2\sigma_\alpha^2=m^2\sigma_\alpha^2>0.$$

This indicates that, given a nondegenerate distribution of α and a constant m, layoffs should be a positive function of the wage after controlling for H.[7] The contrasting signs in expressions (15) and (16) confirm the ambiguity of the transactions cost model with respect to the relationship of the wave level and the likelihood of permanent layoff.

In contrast to the above predictions of the transactions cost model, the costless post-hire negotiations model predicts that quits and layoffs are similarly related, or similarly unrelated, to the level of wages observed at some time prior to job change. Specifically, the costless renegotiations model predicts that quits and permanent layoffs are either both unrelated to the current job wage level or both negatively related to measured wages. To understand this implication, two assumptions of this model should be underscored. First, recall that with costless renegotiation of contracts, both quits and layoffs are determined following the initial period of employment according to the inequality given in equation (11): $m+\eta<\varepsilon$, which implies that a job change occurs only when an alternative job rent share ε dominates current job total realized match value $m+\eta$.

Second, the process generating wages described in (1) must now be reformulated to allow for the dependence of α on the current state of negotiations S,

$$(1')\qquad w=H+\alpha(S)m.$$

This reformulation of the wage-generating process implies that the variance in α now reflects variance in S, or variance over the life of individual job matches, as well as representing differences among individual contracts. Wage variance thus embodies both changes in the specification of individual contracts over time and the heterogeneity of contracts in the sample at any time. The costless negotiations model thus implies that the variance in α should be quite large. This augmented variation in α increases the variance in wages over time, and thus makes prior survey wages a rather tenuous measure of the $m+\eta$ level relevant to the determination of a job change.

We may now more formally clarify the relation of wages and turnover suggested by the costless negotiation model. According to the model assumptions outlined in expressions (11) and (1′), previous survey wages may predict job change if we can use the match-specific components of these wages, αm, to proxy the level of m on the left side of inequality (11). Using previously mentioned independence assumptions, the corre-

[6]It is clear that this expression (14) is really the sum of the variance of αm and minus one times the covariance of m and αm. The covariance of m and αm is derived via the formula, $E(\alpha m^2)-E(\alpha m)E(m)$, which gives the independence of α and m reduces to $\mu_\alpha\sigma_m^2$.

[7]This is the result predicted for layoffs by Parsons' model. According to this model, since all workers invest in equal amounts of m, only the variance of α determines layoffs.

lation of these expressions is[8]

$$(17)\quad \operatorname{corr}(m,\alpha(S)m) = \frac{\mu_\alpha \sigma_m}{\left(\sigma_\alpha^2\sigma_m^2+\mu_\alpha^2\sigma_m^2+\mu_m^2\sigma_\alpha^2\right)^{1/2}} > 0.$$

If this correlation is close to one, then both the quit and layoff probabilities should be negative functions of the wage. Whatever the value of this correlation however, it is clear from (11) that quits and layoffs should be similarly related to the wage.

The actual value of this correlation, and thus the relationship of wages to turnover probabilities in a world of zero or low transactions costs, depends of course on the relative variances in α and m. The costless negotiations model, however, implies that the variance in α should be large reflecting variability over the life of specific contracts as well as differences among separate individual employment agreements at any time. Expression (17) implies that as the variance of α gets large relative to the variance of m, the ability of observed wages to predict job changes declines. That is, as the variance of α gets larger, wages, after controlling for H, become a poorer and poorer proxy for the left side of inequality (11). On the other hand, if the variance of m dominates, quits and layoffs are both negatively related to wages; in this case wages proxy expected productivity rather well.[9]

The crucial difference between the models with and without costly negotiations involves the source and magnitude of variance in α. The model without renegoitation of wages implies that α varies only over individual match contracts. Thus, after controlling for general human capital H, the worker's rent share αm may be estimated at any time via previously observed wages. Since the quit probability negatively relates to both components of αm, the model unambiguously implies that wages are negatively related to quits. The story is different with respect to layoffs however. Parameters α and m cannot be separated empirically, and since α positively affects while m negatively influences the layoff probability, the model cannot predict the relation of layoffs to wages.

The model with costless negotiation, however, assumes that the variance in α reflects variation across matches but also includes variation over time for each individual contract. But referring back to expression (17), it is apparent that this implies that observed wages much less precisely approximate the actual match value which determines quits and layoffs: wages should at best be a rather weak predictor of turnover and, to the extent wages systematically relate to turnover decisions, they should impact quits and layoffs identically.

In summary, the relevant contrasting empirical implications are that, while the costless negotiation model implies a similar weak negative effect of wages on quit and layoff probabilities, the model assuming costly negotiation implies that quits are a strong negative function of the wage and that layoffs are ambiguously related to the wage level. The two models are empirically indistinguishable in the case where both quit and layoff probabilities are observed as negative functions of the wage.

B. *Other Factors Determining Job Change*

This subsection outlines several other factors affecting job change which, although not formally integrated into the model, require some brief consideration. The discussion here also is intended to justify the statistical specification of the next section.

1. *Job Tenure*. The accumulation of job tenure, according to my model, reduces the

[8] I have assumed that the wage is observed prior to the revelation of η. Similar implications hold if η is reflected in the observed wage. Simply define $m^* = m + \eta$, and derive an expression similar to (14) in terms of m^* instead of m. The variance of α still undermines the predictive power of wages. According to the assumptions of the zero transactions cost model, the observed wage may or may not reflect η, thus making the relation of previously observed wages and turnover implied by the zero negotiations cost model even more tenuous.

[9] As a corollary to the current discussion, it can be seen that the usefulness of wages in predicting job change depends on various parameter values even when negotiations are costless.

probability of quit or layoff because longevity increases the possibility that the initial trial period of employment has expired without a job change. Although the model is rather artificial in ignoring the possibility of job change in later periods, the basic selectivity mechanism that drives the attenuating effect of tenure on turnover in this model would operate in more complicated models also. For instance, assuming the wage-offer distributions remain moderately stable over time, those workers with a low probability of finding a better job in the initial period would also be less likely to improve themselves with a quit at a later time. Similarly, if worker productivity is relatively time invariant, an output contribution sufficient to forestall layoff in the first period would imply a lower probability of subsequent layoff.

Although tenure may well predict quits and layoffs in empirical studies, as an explanatory variable tenure alone cannot inform us about the underlying decision processes of workers and firms. Rather, the negative effect of tenure upon mobility only indicates that some aspects of match value to workers and firms tend to persist over time.[10] Tenure alone is of little value in discriminating among various hypotheses concerning the determination of mobility, and for this reason I shall use job tenure only as a general control for various other unmeasured factors in the empirical work.

2. *Short Contract Jobs*. Some jobs are by their inherent nature of short or limited duration. For instance, construction jobs may end at the completion of a certain building or structure. Agricultural employment may be limited by the duration of the harvest. And since the surveys that yielded the data for this study occurred in October and November, many retail sales jobs may have only extended through the holidays. These considerations rationalize the inclusion of various industry dummies in the empirical specification.

3. *Worker-Intentional Short-Term Jobs*. Some young workers may choose temporary employment while deliberating career choices. Thus, we would expect part-time workers or workers in industries such as retail sales to show higher quit rates.

4. *Nonpecuniary Aspects of Job Value*. Workers may value their current job for reasons unrelated to the actual level of wages or earnings. Relationships with other employees and supervisors, the pace of work, or the general work environment may be factors affecting the probability of a quit. The empirical work thus attempts some control for nonpecuniary aspects of work.

5. *Union Membership*. Members of a union should be less at risk to change jobs since unions may better screen workers before hire. Alternatively, fixed costs of joining a union as well as the presence of many benefits contingent upon seniority may attract more serious or committed young workers. Generally, we would expect these factors to augment job permanence among union workers: union membership should be negatively related to quits and permanent layoffs.[11]

III. Empirical Results

This section reports empirical results that tend to support the transactions cost model of mobility. Basically, I find that while wages negatively affect quit decisions, layoffs appear unrelated to observed wage levels. The section commences with a discussion of the data and sample characteristics, however.

[10]A negative tenure and mobility relation would be predicted by various other models describing the accumulation of specific training, models involving heterogeneous worker tastes for the consumption of time constant nonpecuniary job attributes, models describing the accumulation of personal relationship capital on the job, or by models that assume that workers are characterized by varying degrees of taste for job change itself.

[11]Richard Freeman (1976, 1980) has argued that organized channels of communication between workers and managers embodied in union grievance procedures decreases quits by attenuating the need for an "exit-voice" mechanism of communication. That is, union workers don't get mad and quit, but rather, negotiate differences with the firm.

A. *Data*

The data comprises 709 observations derived from the National Longitudinal Surveys of Young Men. Job change behavior was tracked between the 1969 and 1970, and also the 1970 and 1971, contiguous surveys. Wages and other general job information were available for all jobs intervening these surveys from questions asked about recent jobs in the second-year interview. For those observations changing jobs, it could be determined whether job mobility involved a quit or layoff.

Since the sample was intended to describe a population of young men just starting in the labor market, observations were chosen so that no more than two years had elapsed since last school enrollment. As indicated in Table 1, the sample is young and inexperienced. Average education stands at about thirteen years, and, as in the original surveys, nonwhites are overrepresented about two to one. Self-employed and government workers were excluded from the sample.

Table 1 also indicates that about 29 percent of the sample quit their initial survey job for another employer, while only about 14 percent of the sample experienced layoff coincident with job change. According to unreported tabulations by mobility group, quitters and those workers experiencing layoff tend to be about one year younger and less educated than immobile workers. Prior to job change, mobile workers tend to earn less and like their jobs less than constant employer workers. Job changers, in general, also tend to have less job tenure, report less work experience, and are less often employed in financial or union jobs. Mobile workers are more often found in construction, agriculture, retail sales, entertainment services, and part-time jobs.

TABLE 1—DEFINTION OF VARIABLES AND SAMPLE MEANS

Agr = Dummy taking unit value if employed in agriculture	.03
Age = Age	19.84
Att = Attitude towards job[a]	1.70
Bus = Employed in business and repair services	.06
Con = Employed in construction	.13
D70 = Dummy taking unit value for 1970 observations	.49
Ed = Years of schooling	12.74
Ent = Employed in entertainment and recreational services	.01
Exp = Weeks experience	138.68
Finan = Employed in finance	.06
Hlth = Dummy taking unit value if health limits work	.08
Mannon = Employed in nondurable manufacturing	.27
Mine = Employed in mining	.02
Per = Employed in personal services	.02
Prof = Employed in professional and related services	.13
Pt = Part-time status[b]	.17
Ret = Employed in retail sales	.31
Tenure = Months employed	7.62
Tran = Employed in transportation	.10
Whol = Employed in wholesale trade	.05
Wr = Real wage[c]	2.39
Union = Dummy taking unit value if union member	.20

[a]Variable takes on ordered values 1, 2, 3, 4, where 1 indicates likes job very much, and 4 indicates dislikes job very much.
[b]Worked less than 30 hours.
[c]All wages in this study were *CPI* adjusted for the actual month of observation.

B. *Empirical Estimates*

According to the parameter estimates given in Table 2, wages affect the probabilities of quit and permanent layoff differently. The probit estimates indicate a negative and significant effect of wages on quits, but suggest that permanent layoffs are unrelated to the wage level. After controlling for other variables possibly affecting the probability of layoff, the results indicate that in terms of wages, workers experiencing layoff are not distinguished from the majority of the sample who maintained their initial employment during the year interval between surveys. On the other hand, quitters are shown to earn wages far below immobile workers. These results, while consistent with the pattern of wage influence on turnover suggested by the costly negotiations model, run counter to the predictions of the model with costless negotiation of employment contracts. Quits and layoffs are not similarly related to the wage

TABLE 2—QUIT AND LAYOFF PROBIT ESTIMATES[a]

	Quit	Quit	Layoff	Layoff
Wr	−.234[illegible]	−.2220	−.0019	.0298
	(−3.09)	(−2.67)	(−.02)	(.33)
Age	−.026[illegible]	−.0112	−.0210	−.0317
	(−.62)	(−.26)	(−.39)	(−.57)
Ed	−.053[illegible]	−.0552	−.0889	−.1148
	(−1.14)	(−1.13)	(−1.50)	(−1.80)
Exp	.001	.0008	.0020	.0021
	(1.28)	(.93)	(1.75)	(1.67)
Tenure	−.020[illegible]	−.0221	−.0786	−.0746
	(−2.20)	(−2.33)	(−5.34)	(−4.92)
Union	−.474[illegible]	−.4540	−.1817	−.0761
	(−2.90)	(−2.69)	(−.96)	(−3.7)
Att	.3995	.4002	.4305	.4776
	(4.96)	(4.87)	(4.48)	(4.68)
Pt	.938[illegible]	.8870	1.023	.8734
	(5.90)	(5.41)	(4.96)	(4.03)
Hlth	−.040[illegible]	.0017	.4042	.3853
	(−.18)	(.01)	(1.63)	(1.47)
Nwht	−.048[illegible]	−.0306	.0970	.1389
	(−.32)	(−.20)	(.54)	(.73)
D70	−.026[illegible]	−.0255	.0233	−.0570
	(1.04)	(−.20)	(.15)	(−.34)
Agr	–	.2965	–	.9066
		(.68)		(1.80)
Mine	–	.5432	–	–
		(1.25)		
Con	–	.5569	–	.8054
		(2.39)		(3.17)
Mannon	–	.2028	–	−.2339
		(1.08)		(−.88)
Tran	–	.2689	–	−.3274
		(.93)		(−.70)
Whol	–	.1130	–	.4189
		(.31)		(1.02)
Ret	–	.3912	–	.4137
		(2.15)		(1.84)
Finan	–	−.3876	–	–
		(−1.00)		
Bus	–	.1692	–	.7400
		(.51)		(2.06)
Per	–	−.1865	–	.4423
		(−.31)		(.67)
Ent	–	1.209	–	.7972
		(1.58)		(.89)
Prof	–	.1651	–	.5750
		(.60)		(1.66)
Constant	.58[illegible]7	.1077	−.0442	.0645
	(1.04[illegible]	(.17)	(−.06)	(.07)
Log-likelihood	−321.55	−314.38	−193.18	−182.18
Sample Size	613	613	503	503

[a]The t-statistics are shown in parentheses. Durable manufacturers are the excluded industry group.

level as predicted by the marriage model of employment contracts.[12]

Age, education, and experience were included to control for the general human capital component of wages H. The overall impact of these variables on turnover appears nil except for the layoff probits where education shows a slight inhibiting effect and experience exhibits a small positive influence.

Tenure was included to control for unobserved worker and job match heterogeneity. Tenure exercises a negative and significant effect on both quits and layoffs; a pattern consistent with my model but also suggested by competing explanations of turnover. It is notable also that tenure shows a far larger impact on layoffs than quits.[13]

Union status acts as a significant deterrent to quits but shows no significant effect on layoffs. These results for quitters are generally consistent with the notions that unions attract more serious workers or that unions attenuate the exit-voice phenomenon.

Attitude toward the job significantly affects both quits and permanent layoffs. Workers who do not like their jobs are both more likely to quit and experience layoff. This variable was included primarily to control for the impact of nonpecuniary dimensions of job value on quits. The similar effect of attitude on layoffs, however, suggests that this variable is probably picking up on negative job characteristics associated with temporary jobs.

Part-time status was included to control for any worker or firm-motivated temporary job phenomenon. Part-time status manifests a positive and significant effect on both quits and layoffs. This generally positive influence is consistent with the notion that part-time status proxies worker seriousness or job permanence.

The health status variable is generally insignificant but exhibits much larger effects on layoffs than quits. Indicators of race or sample year show no significant effects on turnover.

Industry dummies were included to measure short-term job phenomenon and also to control for any possible compensating wage differential effects. Inclusion of these variables does not alter the basic pattern of the previously discussed results except that the wage coefficient in the layoff probit becomes positive. Workers in construction and retail sales are more likely to quit or experience layoff. Agricultural employees along with business and professional services workers are more likely to be laid off. This pattern of results is also roughly in accordance with the discussion of the previous section.

[12] Some mention should be made concerning the likely direction of bias resulting from the difficulties in distinguishing quits from permanent layoffs with the present data. Although the categorized questionnaire responses are unambiguous, two types of misreporting are possible. First, workers may feel some stigma attached to experiencing a layoff and thus report a quit. Second, while on temporary layoff, a worker may search and then in effect quit his old job but honestly report layoff as the reason for leaving. In as much as temporary layoff behavior is generally associated with more experienced workers receiving unemployment insurance subsidies, the youthfulness of the sample probably minimizes the impact of the second source of bias. More generally however, since the effect of either form of misreporting is to make reported quits and layoffs look similar, the empirically supported conclusion of this paper, that quits and permanent layoffs are different, is unaffected. Note also that layoffs and quit observations are excluded from the quit and layoff probits, respectively. Software limitations precluded estimation of a model simultaneously explaining both mobility components. This, of course, precludes directly performing a classical statistical test of quit-layoff differences. Rather, I contrast quits and layoffs by comparison to a common group of immobile workers. Similar probit specifications were run separately comparing quitters and workers experiencing layoff to the rest of the sample. For example, I defined the dichotomy in terms of quitters vs. layoffers and stayers. Those probits yielded wage coefficients indicating even a greater difference between quits and layoffs. This specification had the unfortunate effect of attenuating and compromising the interpretation of many other coefficients where the related explanatory variable was expected to similarly impact quits and layoffs, however.

[13] Wage and tenure interactions were added to the probit specifications to determine whether the wage effects were constant with respect to time on the job. None of the interaction coefficients were even near significance at commonly accepted levels, while the layoff and quit probit wage coefficients from these extended specifications remained similar in terms of size and significance to those parameters in Table 2.

In summary, the empirical results reported in Table 2 appear consistent with the earlier discussion. In particular, the wage coefficients suggest that quits and permanent layoffs are different in a way consistent with the predictions of the model of job matching with costly contract renegotiation.

The regression results reported in the first column of Table 3 offer some further evidence concerning the wage implications of turnover. The regression coefficients reported in the second column of Table 3 also serve to allay concerns about possible omitted variable bias effects of unobserved ability on the wage coefficients reported in the quit and layoff probits above.

The first column of Table 3 reports estimated coefficients from a wage change regression equation commonly used in studies of mobility and earnings (Ann Bartel, 1979; Matthew Black, 1980; Steven Sandell, 1977). The dependent variable is the change in real log-wages $\Delta \ln W$. For mobile workers the change in log-wages represents the difference between the old and new job hourly earnings. For immobile workers the wages were measured at the first and second survey time. All sample members were at risk to change jobs in the year interval between surveys. All wages were *CPI* deflated according to their month of observation to assure comparability across time and mobility groups

The education, experience, race, and survey year variables were included to control for life cycle earnings profile effects and time differences. The variable ΔExp represents the change in weeks experience between wage observations.

The turnover variables reported in the first column indicate that, while quitters unambiguously gain from mobility, workers experiencing layoff show borderline significant wage losses. These results are hard to reconcile with the notion that layoffs are really only anticipated quits. Here, as before, quits and permanent layoffs appear different in a manner consistent with the costly negotiations theory of job matching. Apparently, quitters transfer to higher paying jobs while individuals who do not quit but experience layoff are forced often to settle for lower wage employments. Quits and permanent layoffs, at least in this young sample, again appear different.[14]

TABLE 3—WAGE GROWTH AND WAGE LEVEL REGRESSIONS[a]

	$\Delta \ln W$	$\ln W73$
Quit	.0743	−.0259
	(2.55)	(−.67)
Layoff	−.0666	.0218
	(−1.65)	(.42)
Δ *Exp*	.0039	–
	(4.15)	
Exp	-2.1×10^{-6}	–
	(−.004)	
Exp × *Exp*	-2.8×10^{-7}	–
	(−.26)	
Ed	−.0012	.0940
	(−.20)	
NWht	−.0300	−.2110
	(−1.00)	(−4.87)
D70	.0279	−.1049
	(1.13)	(−3.09)
Exp 73	–	−.0702
		(−1.01)
Exp 73 × *Exp* 73	–	.0051
		(1.54)
Constant	−.0756	.4927
	(−.80)	(1.32)
R^2	.06	.26
F	5.18	28.48
Sample Size	709	566

[a]The t-statistics are shown in parentheses; $\ln W73$ represents the nominal wage at the 1973 survey time.

[14]These wage growth results on quits are quite strong compared to other mobility studies using young samples (see Joseph Antos and Wesley Mellow, 1979; Bartel, 1979). This disparity is explained by my data organization that serves to eliminate a serious misspecification present in many other studies. Previous studies have attempted to statistically relate wage profiles and mobility decisions by measuring wage growth between surveys while defining a job change as voluntary or involuntary according to whether interviewees report a quit or layoff from the first survey job. However, since an initial job change may be followed by another quit or layoff prior to the second survey, growth or decline in wages between surveys may only tenuously relate to the explanation given for leaving the initial survey employer. My methodology eliminates this form of specification error associated with many previous studies. A comparison of regressions using the correct interjob wage differences with regressions using intersurvey wage differences revealed that the mobility coefficients were underestimated by as much as 50 percent. It is probably reasonable to conclude that the extent of the bias in using survey wages rather than specific job wages varies in-

Finally, the coefficient estimates in the second column of Table 3 derive from the regression of 1973 survey log-wage on various typical human capital variables and earlier mobility status. The results indicate that, by 1973, mobile workers in the earlier interval exhibit wage levels comparable to earlier immobile workers. These results are not intended to measure the effects of mobility on earnings, this they cannot do.[15] Rather, the insignificant mobility coefficients indicate that permanent earning ability differences do not explain the earlier reported effects of wages on quits and permanent layoffs. This wage regression suggests that within a few years of the original mobility interval, mobile and immobile workers appear similar with respect to earnings. The initial wage differences among the mobility groups appear to have eroded. These results are consistent with the work of other authors (Bartel, 1980; Wesley Mellow, 1978), but more importantly in the present context strengthen my conclusions derived from the previous quit and layoff probits. Specifically, these results suggest that the effects of wages on turnover reported in Table 2 are not biased by the omission of ability from the probit regressions. Mobile workers are not systematically less able. Quits and permanent layoffs are differently related to current job wage levels for reasons independent of unobserved ability.

versely with the age of the sample, and varies directly with the time elapsed between surveys. Older workers tend to be less mobile so the likelihood of multiple intrasurvey moves is smaller, and the probability of multiple moves increases with the time between surveys. Note also that any effect of unmeasured earning ability on these coefficients is eliminated by differencing wages.

[15]For example, quitters may gain from mobility but in the process only catch up to immobile workers. Thus, these wage level regressions should not be used to measure mobility gains or losses. Also note that some observations were lost following the 1971 interview. Ancillary sample attrition probits indicated that the probability of attrition was unrelated to earlier mobility status, however.

IV. Conclusion

This paper outlined a model of job matching appropriate to a population of inexperienced young men. The empirical results indicated that quits and permanent layoffs are different. More specifically, these results tended to confirm the model of job matching with costly contract renegotiation. It should be stressed, however, that these conclusions need not generalize to older populations. Productivity uncertainty is an important assumption of my model. To the extent this uncertainty is attenuated as workers age and become "known commodities," the conclusions of this paper might need to be modified.

Finally, an interesting implication of the model concerning the relation of productivity to observed wages should be recognized. Conditional upon the acceptance of the costly contract renegotiations model, the wage coefficient in the layoff probit suggests the extent to which the variance in actual productivity m, or the variance in the rent-sharing parameter α, dominates the variance in the match-specific component of the wage αm. The theoretical discussion suggested that dominance of the m variance implied a negative layoff and wage relation, while dominance of the variance in α implied a positive association between the wage level and the layoff probability. The empirical results generally showed a zero wage coefficient in the layoff probits. This suggests a roughly equal role for rent sharing and productivity in the determination of wages. It thus appears, at least for very young workers, that firms may capture some match rent. This implication, while at odds with the usual neoclassical market model, is perhaps not surprising in a real market setting with heterogeneous agents and costly acquisition of information.

REFERENCES

Antos, Joseph R. and Mellow, Wesley, "The Youth Labor Market: A Dynamic Overview," BLS Staff Paper 11, U.S. Department of Labor, July 1979.

Bartel, Ann P., "Earnings Growth on the Job and Between Jobs," *Economic Inquiry*, January 1980, *18*, 123–37.

______, "The Migration Decision: What Role Does Job Mobility Play?," *American Economic Review*, December 1979, *69*, 775–86.

Becker, Gary S., Landes, Elisabeth M. and Michael, Robert T., "An Economic Analysis of Marital Instability," *Journal of Political Economy*, December 1977, *85*, 1141–87.

Black, Matthew, "Pecuniary Implications of On-The-Job Search and Quit Activity," *Review of Economics and Statistics*, May 1980, *62*, 222–29.

Freeman, Richard B., "Individual Mobility and Union Voice in the Labor Market," *American Economic Review Proceedings*, May 1976, *66*, 361–68.

______, "The Exit-Voice Tradeoff in the Labor Market: Unionism, Job Tenure, Quits and Separations," *Quarterly Journal of Economics*, June 1980, *94*, 363–73.

Hashimoto, Masanori, "Bonus Payments, On-The-Job Training, and Lifetime Employment in Japan," *Journal of Political Economy*, October 1979, *87*, 1086–104.

______, "Firm-Specific Capital as a Shared Investment," *American Economic Review*, June 1981, *71*, 475–82.

Jovanovic, Boyan, "Job Matching and the Turnover," *Journal of Political Economy*, October 1979, *87*, 972–90.

Lippman, Steven A. and McCall, John J., (1976a) "The Economics of Job Search: A Survey: Part I," *Economic Inquiry*, June 1976, *14*, 155–89.

______ **and** ______, (1976b) "The Economics of Job Search: A Survey," *Economic Inquiry*, September 1976, *14*, 347–68.

Mellow, Wesley, "Equilibration in the Labor Market," *Southern Economic Journal*, July 1978, *45*, 192–204.

Mortensen, Dale T., "Specific Capital and Labor Turnover," *Bell Journal of Economics*, Autumn 1978, *9*, 572–86.

Parsons, Donald D., "Specific Human Capital: An Application to Quit Rates and Layoff Rates," *Journal of Political Economy*, November 1972, *80*, 1120–43.

Sandell, Steven H., "Women and the Economics of Family Migration," *Review of Economics and Statistics*, November 1977, *59*, 406–14.

Center for Human Resource Research, *The National Longitudinal Surveys Handbook*, Columbus: College of Administrative Science, The Ohio State University, 1981.

The Demand for and Supply of Births: Fertility and its Life Cycle Consequences

By MARK R. ROSENZWEIG AND T. PAUL SCHULTZ*

Economists and other social scientists have studied since Malthus the determinants and consequences of fertility, but without evolving a consistent framework. Two lines of fertility-related research have emerged in the economics literature. The first, concerned chiefly with the determinants of fertility, emphasizes the preferences of parents and treates the number of children born to a household as wholly demand-determined. Children are viewed as one manifestation of the set of household decisions constrained by costs and income (Schultz, 1969; Robert Willis, 1973; Rosenzweig and Robert Evenson, 1977; William Butz and Michael Ward, 1979). Within this tradition, estimates of the "effects" of fertility variation on household behavior are obtained using simultaneous equation techniques (for example, Marc Nerlove and Schultz, 1970; Belton Fleisher and George Rhodes, 1979); however, since both fertility and other household goods are viewed as jointly "demanded" by the household, identification appears arbitrary, in which case such estimates provide little information about the consequences of fertility change.

A second line of research assumes that births are wholly supply-determined, and treats fertility as purely random and exogenous. The number and ages of children in a household do not enter the household's preference function and/or are treated as exogenous (unexplained) determinants of or constraints on household allocation decisions, in particular, the supply of female time to the market (James Heckman and Thomas MaCurdy, 1980; Jerry Hausman and Paul Ruud, 1984) and household consumption expenditures (Robert Pollak and Terence Wales, 1980; Angus Deaton, 1982).[1]

The coexistence of these mutually inconsistent approaches to the study of fertility-related behavior may reflect the special nature of fertility. Unlike the consumption by households of television sets or food, the level of fertility is determined by the allocation of resources to limit the biologically determined production of births—fertility supply. Given that such control is costly and imperfect, and the biological capacity to bear children (fecundity) is stochastic and mostly unaffected by choice behavior, the number of children born to a couple (society) may not exactly correspond to either the couple's (societal) expectations of or preferences for (given costless control) its family (population) size. Moreover, an individual couple may learn about its own specific supply constraints over time; planned resource allocations to control births or to earn income, therefore, may be altered by fertility experience both directly and indirectly as it influences supply expectations.

*Professors of Economics, University of Minnesota-Twin Cities, Minneapolis, MN 55455, and Yale University, New Haven, CT 06520, respectively. This research was supported in part by grants from NIH, Center for Population Research, HD-12172 and the Hewlett Foundation to the Economic Demography Program at Yale. We have benefited from the comments of the referees, Jere Behrman, Julie DaVanzo, James Heckman, Evelyn Lehrer, Daniel McFadden, Norman Ryder and members of the Yale Labor and Population Workshop and the MIT Econometrics Workshop.

[1]Exceptions to these traditions are Heckman and Willis (1975), and Rosenzweig and Kenneth Wolpin (1980). Heckman and Willis construct a dynamic model of contraceptive choice incorporating stochastic fertility. However, their econometric application does not consider the behavioral implications of their model (see Section II, below). Rosenzweig and Wolpin exploit the incidence of twins to estimate the exogenous impact of an additional birth on life cycle labor supply, but in the context of a perfect foresight, deterministic model.

Fertility within a household is determined by the dynamic interaction between its supply of and demand for births, and variations in births across households reflect exogenous intercouple variation in both the supply of births and prices, income, and preferences for children, or demand.[2] Economists have recognized the joint relevance of biological and behavioral factors in determining fertility under a regime of costly fertility control (for example, Richard Easterlin, Pollak, and Michael Wachter, 1980), but this perception has not been suitably incorporated into the empirical study of fertility.[3] In particular, this insight has not been employed in estimating the effectiveness of contraceptive methods, or in estimating the effects of fertility variation on labor supply behavior, on earnings, or on the demand for fertility control. When there is heterogeneity in both the biological supply of births and in preferences for family size, neither of which are directly observed by the researcher, a new analytical approach is needed.

In this paper we develop and implement a methodology for identifying empirically the supply function for births in order to assess the consequences of exogenous variation in the supply of births for fertility, for couples' choice of contraceptive techniques, and for the life cycle labor supply and earnings of married women. In Section II we describe an illustrative dynamic optimization model of contraceptive choice and labor supply incorporating uncertain fertility and persistent heterogeneity in preferences and fecundity. Section III describes the strategy for obtaining estimates of the biological components of fertility and the efficacy of contraceptive techniques, given household optimizing behavior and heterogeneity in fertility supply and preferences. In Sections IV, V, and VI, the methodology is applied to monthly longitudinal data on contraceptive use, fertility, and female labor supply in the United States.

The empirical results indicate that, because of the costliness of contraception, more than 10 percent of the cross-sectional variation in the number of live births in the United States is due to interhousehold variation in the exogenous supply of births, with couples offsetting 25 percent of the underlying variation in fertility supply through their choice of contraceptives. Such compensatory behavior appears to lead to significant underestimates of the efficacy of contraceptive methods, unless heterogeneity is taken into account. Biologically determined fertility supply variation appears also to affect significantly married women's labor supply and earnings. Moreover, use of actual fertility as a proxy for fertility supply, as in prior studies, overestimates the true life cycle labor supply response of married women to exogenous variations in fertility associated with imperfect and costly fertility control, but underestimates their consequences for female earnings and for the life cycle adjustment in couples' choice of contraceptive methods.

[2] It is widely believed that fecundity varies substantially across couples but that this reproductive potential is not greatly affected by contemporary socioeconomic conditions. Although acute malnutrition and certain diseases may induce premature sterility and thereby influence the distribution of fecundity in low-income populations, fecundity is essentially exogenous in a high-income population today, such as that of the United States in 1970–75 (John Bongaarts and Jane Menken, 1983). In the United States, contraceptive behavior is naturally seen as the important source of behavioral variation in fertility. Nonetheless, an integrated framework for analyzing both the supply of and demand for births is needed if the consequences of specific sources of fertility variation are to be assessed.

[3] In the social science literature, Rodolfo Bulatao and Ronald Lee (1983) distinguish among the determinants of fertility supply and demand and the costs of fertility regulation without attention to the underlying sources of stochastic variability and to how covariation between them might complicate the task of statistically identifying the effects of exogenous biological supply from those of endogenous behavioral demand factors. Using an analogous framework, Bongaarts and Robert Potter (1983) examine the proximate determinants of fertility. They also ignore the manner in which the biological proximate factors determining fertility are self-selected, and thus are a potential source of bias in conventional calculations of contraceptive use-effectiveness from nonexperimental populations represented in survey data. Easterlin and Eileen Crimmins (1982) demonstrate some of the difficulties in estimating elements of a supply and demand model, but they ignore the stochastic interplay of demand and supply, the involuntary sources of biological variation in fecundity, and the individual optimization process.

I. A Dynamic Model of Fertility Control and Labor Supply with Exogenous Fecundity Variation

Assume that the number of births N_{ij} of couple j in period i in which the mother is age $i+a$ is a random variable that can be reduced by the use of resources to control fertility, Z_{ij}, such that

$$(1) \quad N_{ij} = \mu_j + \varepsilon_{ij} + \gamma(a+i)_j - \beta Z_{ij},$$

where μ_j is a time invariant fertility component specific to couple j, ε_{ij} is an independently distributed, serially uncorrelated disturbance, and γ and β are biological-technical parameters to be estimated.[4]

Equation (1) describes the technology of birth production, or what is called hereafter the reproduction function. The potential number of births or fecundity is the sum of the first three exogenous terms and may be referred to as the "supply" of births; since the number of births actually produced (fertility) depends as well on the use of fertility control, realized births N_{ij} depend on both the preferences of parents (demand) and the supply of births.

Assume that parents in each period maximize the expected value of an intertemporally separable utility function that has as arguments the number of children born in the period, the cumulated number of children M_i (where $M_i = M_{i-1} + N_i$), a consumption good X_i, and leisure L_i of the mother. The parents' problem at period s is described by

$$(2) \quad \max E\left[\sum_{i=s}^{t} \delta^{i-1} U(N_i, M_i, X_i, L_i; \tau)\right]$$

subject to (1) and to a within-period income constraint,

$$(3) \quad W_i(T-L_i) + V_i = pZ_i + X_i + cM_i,$$

where the couple subscript j is suppressed, W is the wage of the mother, T is her total available time, V is income of the husband, p is the cost of fertility control, c the per period cost of a child, τ is a time-invariant, couple-specific random preference variable, where $E(\tau_j\tau_k) = \sigma_\tau^2$ if $j=k$ and zero otherwise and $E(\tau_j\mu_k) = 0$ for all j and k.[5] This latter assumption presumes that fecundity is not influenced by behavior (see fn. 2). Parents are thus heterogenous in preferences (τ_j) and in the technologically persistent potential supply of births (μ_j).

The household chooses whether to use fertility control ($Z=1$) in each period, and whether or not the mother is to work ($L=0$), based on the information set Ω at the beginning of the period.[6] At the onset of period i, then, parents know the outcomes of their past decisions, the technology of birth production β, the wage rate, prices, and their preferences; they thus know M_{i-1} but do not know the fertility they will experience in period i. They *may* know that component of the random fertility "draw" which is time-invariant (μ).

In such dynamic problems it is not generally feasible to derive analytically decision rules for L_i and Z_i in any period (Kenneth Wolpin, 1984). However, comparative statics can be performed for the final-period decisions, when parents need not take into account the impact of their current decisions on future decisions. To see how variations in "fertility" are related to decisions concerning both contraception and labor supply as functions of technology, preferences, and costs, assume that the utility function in each period

[4] It is assumed that $E(\mu_j\mu_k) = 0$ if $j \neq k$, $= \sigma_\mu^2$ if $j=k$; $E(\mu_j\varepsilon_{ij}) = 0$, $E(\varepsilon_{ij}\varepsilon_{kj}) = 0$ if $i \neq k$, $= \sigma_\varepsilon^2$ if $i=k$. We abstract in (1) from the role of sexual intercourse and other behavioral determinants (diet, exercise) of reproductive potential for simplicity. Intercourse behavior and smoking by the mother are incorporated in the empirical analysis below.

[5] We abstract from the dependence of wage rates on prior labor force experience. In the empirical analysis below we estimate the cumulative effects on female earnings due to the adjustments in labor force experience resulting from exogenous variation in the supply of births.

[6] In the empirical work below, we consider a variety of methods and the choice among them.

is linear quadratic:

$$(4)\quad U_i = \alpha_1 N_i - 0.5\alpha_2 N_i^2 + \alpha_3 M_i - 0.5\alpha_4 M_i^2 + \alpha_5(\tau) X_i - 0.5\alpha_6 X_i^2 + \alpha_7 L_i - 0.5\alpha_8 L_i^2 + \alpha_9 L_i M_i,$$

where $\alpha_1, \dots, \alpha_8 > 0$, to reflect diminishing marginal utility of all commodities. Note that mother's leisure and fertility are allowed to interact ($\alpha_9 \neq 0$), reflecting the presumption of the household production literature that these two variables are not independent. We abstract from other interactions in (4) for simplicity, as is common in labor supply and fertility models. Preference parameters are constant across all periods, but preferences for the consumption good X vary across households, such that $\alpha_5(\tau)' > 0$, for all i.[7]

Consider first for illustrative purposes the decision whether to use fertility control in the last period, t. It can be shown that the difference in expected utilities, J_t^Z, associated with the use and nonuse of contraception is

$$(5)\quad J_t^Z = E_t(U_t | Z_t = 1;\ \Omega_t) - E_t(U_t | Z_t = 0;\ \Omega_t).$$

The relationship between variations in prior fertility M_{t-1} and the probability of using birth control, whether J_t^Z is greater than or less than zero, depends on the information held by parents and on whether the variation in cumulated births is due to preferences or technology. To see this, consider two cases: first, assume that parents know the persistent, family-specific part of fecundity, and thus can distinguish random from systematic fertility outcomes, that is, $\Omega_t = \{\mu,\ \beta,\ \gamma,\ M_{t-1},\ p,\ c,\ W\}$. The effect of a purely random increase in the stock of children in period $t-1$ on J_t^Z is

$$(6)\quad \partial J_t^Z / \partial \varepsilon_{t-1 | \Omega = \{\mu\}} = \alpha_4 \beta - \alpha_6 c(p - \beta c)$$

which, due to diminishing marginal utility, is positive as long as the "savings" from an averted birth βc exceed the cost of using fertility control p. An unanticipated random birth in period $t-1$ increases the likelihood that birth control is used in period t; this effect is stronger the lower the cost of contraception p and the greater its effectiveness β.

The effect of variation in that part of fecundity which is persistent (and known) on the probability of using contraception in period t is given by

$$(7)\quad \partial J_t^Z / \partial \mu |_{\Omega = \{\mu\}} = \alpha_4 \beta (1 + dM_{t-1}/d\mu) + \alpha_2 \beta + \alpha_6 c(p - \beta c) = \partial J_t^Z / \partial \varepsilon_{t-1} + \beta(\alpha_2 + \alpha_4 dM_{t-2}/d\mu).$$

Expression (7) describes how contraceptive use varies across households with differing fecundity, when fecundity is known by the households. Households containing more fecund women are more likely to use fertility control because: (i) they will have more children at the beginning of the period; and (ii) they can expect more children in the future period. As long as more fecund women also have more realized births cumulatively (i.e., $dM_{t-2}/d\mu > 0$), there will be a positive association between individually persistent fecundity and use of contraceptives.

In the case where households treat "excess" births (net of control) as purely random even though fecundity persistence exists, the association between J_t^Z and fecundity is

$$(8)\quad \partial J_t^Z / \partial \mu = \alpha_4 \beta (1 + dM_{t-1}/d\mu) + \alpha_6 c(p - \beta c) = dJ_t^Z / d\varepsilon_{t-1} + \alpha_4 \beta (dM_{t-1}/d\mu).$$

More-fecund women will still be more likely

[7]Other preference parameters may also vary across the population; we show that heterogeneity in one parameter is sufficient to generate biases in estimates of the effects of fertility supply variation on allocative behavior based on associations between household choices and actual fertility.

than less-fecund women to use contraception. As long as higher fecundity increases realized births, lack of knowledge about individually persistent fecundity decreases, but does not eliminate, the positive association between the likelihood of control and fecundity.

Women thus modify their selection of contraceptives on the basis of their realized births. Given the persistence of fecundity over time, estimates of the effect of contraceptive use on fertility, based on samples of women who have already had some fertility experience, will be biased, if heterogeneity is not taken into account. In particular, if, as our framework suggests, more-fecund women use contraceptives more often or more efficaciously, contraceptive effectiveness will be underestimated.

If there is also heterogeneity in preferences, the direct association between prior fertility and current contraceptive use, as estimated in Robert Michael (1973), Michael and Willis (1975), and Daniel McFadden (1975), will not indicate how fecundity affects contraceptive choice.[8] Thus the direction and magnitude of the bias in studies of contraceptive effectiveness cannot be inferred from these studies, since the distribution of actual cumulative fertility across women will reflect both variation in preferences, τ, and in fecundity, μ. The slope of the association between cumulative fertility and current values of the utility differential associated with contraceptive choice in a population heterogenous in fecundity and in preferences consists of three terms:

$$(9)\quad dJ_t^Z/dM_{t-1} = \partial J_t^Z/\partial \varepsilon_{t-1} + [dM_{t-1}/d\tau]^{-1}(\beta c - p)\alpha_6 + \beta\alpha_2 + \alpha_4(dM_{t-1}/d\mu).$$

Women with higher cumulative fertility at the start of period t are likely to be more fecund, and thus will be more likely to use contraception in period t, for given preferences; however, higher-fertility women are also more likely to prefer children relative to other goods X, for given fecundity ($dM_{t-1}/d\tau < 0$), and thus will be less likely to contracept in period t, since they will derive less utility from the savings generated ($\beta c - p$) by contraception. The estimated relationship between prior fertility and current contraception is likely to underestimate both the effect on contraceptive practice of an unanticipated change in fertility and the effect of fecundity variation on contraceptive choice; clearly, the estimated relationship between lagged fertility and contraceptive use does not correspond to either of these analytically distinct effects.

Similar considerations apply to associations between past fertility and the woman's current labor supply decisions, as estimated in most female labor supply studies (James Smith, 1980). The model indicates that women at the start of period t will compare expected utilities in period t with respect to the labor force participation decision L_t; that is,

$$(10)\quad J_t^L = E_t(U_t|L_t=1;\ \Omega_t) - E_t(U_t|L_t=0;\ \Omega_t).$$

The effect on the expected utility differential J_t^L of an unanticipated increase in the number of births in period $t-1$ is given by

$$(11)\quad \partial J_t^L/\partial \varepsilon_{\tau-1} = \alpha_9 - \alpha_6 Wc.$$

The sign of (11) is ambiguous. Unanticipated "excess" fertility increases the marginal utility of the X-good relative to that of children and thus increases the returns to work. However, if children and the mother's home time are complements ($\alpha_9 > 0$), the returns to remaining at home also increase. The relationship between labor force participation and fecundity is given by

$$(12)\quad \partial J_t^L/\partial \mu = (\alpha_9 - \alpha_6 Wc)(1 + dM_{t-1}/d\mu),$$

which again depends on whether or not more-fecund women have higher cumulative

[8] In each of these studies of contraceptive choice, the samples are stratified by the number of children born.

fertility at any age. However, the estimated association between prior fertility and the expected utility differential associated with the current labor supply decision in a population characterized by latent variation in both fecundity and preferences (for X) is

$$(13) \quad \frac{dJ_t^L}{\delta M_{t-1}} = \frac{\partial J_t^L}{\partial \mu} - \left[\frac{dM_{t-1}}{d\tau}\right] W\alpha_6.$$

The dynamic response of mother's leisure (or, conversely, market labor supply) to exogenous variation in fertility due to persistent and random fecundity is confounded by preferences. Since families with higher preferences for consumption good X are likely to have lower cumulative fertility at any age, lower-fertility couples will be more likely to prefer increased X or market income; the labor supply "response" to prior fertility "shocks" measured by (13) rather than by (11) or (12) will be biased upwards in absolute value.

II. Estimating Dynamic Responses to Fertility Supply Variation

The simple dynamic formulation of the contraceptive choice and labor supply problems in the presence of costly fertility control and exogenous variation in fecundity implies that when a couple realizes an unanticipated birth, this "supply shock" is likely to reduce fertility demand, and be manifested (i) in the subsequent adoption of a contraceptive regime that is more effective than it would otherwise have been, and (ii) in a subsequent reduction in labor supply, if home time and children are complements. If the source of the positive shock in fertility is a persistent individual effect, called here "fecundity," then the tendencies to use more effective contraceptives and to restrict labor supply are strengthened, for persistence implies that the couple will already have more births before the current period and the couple may appreciate that they can expect more births in subsequent periods, other factors being equal. This larger number of births cumulated before the current period strengthens the motivation to restrict further births, whether or not the couple can identify its individual fecundity. Thus, the absolute size of the behavioral response to innovations from fecundity (μ) should (i) be larger than the response to random shocks, (ε), and (ii) increase as the couple ages, since older couples will have cumulated more fertility experience. Older couples are also more likely to be able to distinguish random from couple-specific persistent errors, compared to younger couples.

As noted, since realized births may reflect both persistent and unobserved preferences and biological supply factors, tests of propositions about the consequences of fertility supply variation require the isolation of the persistent and exogenous supply components of fertility from their demand components. As can be seen from the reproduction function (1), knowledge of β, the effect of contraception on fertility (or a vector of β's if there are multiple contraceptive devices), combined with information on contraceptive use and realized births, would enable estimation of the μ and ε_{ij} as in any period i, $\mu_j + \varepsilon_{ij} = N_{ij} + \beta Z_{ij} - \gamma(a+i)_j$.

The dynamic interactions between the supply of and demand for births implied by the model, combined with heterogeneity in couple-specific supply propensities imply, however, that information on the effectiveness of contraceptives based on observed associations between conception rates and contraceptive use will be misleading. Heckman and Willis have shown that persistence in couple-specific fecundity yields biased estimates of the β's, and estimate that the couple-specific intertemporal correlation between monthly conception rates is over .5. Their maximum likelihood procedures, however, do not take into account relationships between contraception and fecundity; that is, couples discontinuing use of a particular contraceptive or switching contraceptives in response to their fertility (or nonfertility) experience. Our model implies that couples using a particular technique (after some fertility experience) will not be randomly (self-) selected with respect to fecundity.

Our strategy for obtaining consistent estimates of the β's, and thus of the exogenous supply components of fertility, is to use

time-aggregated information on conceptions, pregnancies, and contraceptive use in an instrumental variable procedure.[9] We estimate the conception rate for a couple over some period as a function of the fractions of that aggregated period the couple used different types of contraceptives. If equation (1) is aggregated over S periods and fertility control Z is used in f of these periods, then the time-aggregated version of (1) is

$$(1') \qquad n_j = \mu_j + \sum_{i=1}^{S} \varepsilon_{ij} + \gamma A_j - \beta F_j,$$

where $F_j = f_j/S =$ the fraction of the aggregate period of exposure in which birth control is used, $n = \Sigma_{i=1}^{S} N_{ij}/S =$ birth or conception rate, and $A =$ average age in the period.

The model implies that F_j will be correlated with both the ε_{ij} and with μ_j (serial correlation is not necessary), as couples will, during the period, exercise control or not in response to within-period supply shocks. Instrumental variable estimates of the self-selected inputs would provide, however, consistent estimates of β, and thus of μ_j and the aggregated ε_{ij}. Because F_j reflects the couple's demand for children, it is a function of preferences, prices (p, c, W_j), and income (V_j). As long as these variables are orthogonal to fecundity, the usual set of fertility demand variables may thus serve as instruments for F_j and permit identification of the supply technology.[10]

III. Data and Estimation of the Reproduction Technology: The Supply of Births

Data are drawn from a longitudinal sample of women from the 1970 *National Fertility Survey* (*NFS*) who were reinterviewed in 1975. The 1970 sample was a national representative sample of 6752 ever-married women born since July 1, 1925. Two thousand three hundred and sixty-one women of the original 1970 sample were reinterviewed in 1975. Their selection was based on being in intact first marriages (for both spouses), where the wife's age at marriage was less than 25, and the duration of the marriage was less than twenty years at the time of the first interview in 1970. Only white couples were included in the 1970–75 panel. A description of the 1970 *NFS* and an outline of the design of the 1975 resurvey are presented in Charles Westoff and Norman Ryder (1977).

The survey provides a month-by-month calendar of contraceptive use, by technique; pregnancies, pregnancy outcomes, intercourse behavior of the couple (abstinence or not), and labor market behavior by the wife from the 1970 to the interview data in 1975, as well as the usual socioeconomic information in both 1970 and 1975. The residential location of the couple in 1970, which is assumed relevant for reproduction in the period, is also provided. Based on this latter information, a series of variables were appended to the micro data to describe the state or SMSA in which each couple resides,

[9] An alternative strategy would be to parameterize the distribution of unobservables and to estimate jointly the reproductive technology and preference parameters associated with the model of Section II. However, this estimation *cum* dynamic programming approach is extremely expensive and would necessitate severe simplifications with regard to functional forms and especially with regard to the number of control variables (Wolpin). Instead, we employ a method which enables estimation of a more complex technology, involving multiple methods of control. We estimate reduced-form effects of persistent and random fertility supply shocks, reconstructed from the estimates of the reproduction technology, on resource allocations over the life cycle. These are based on estimating equations in which only variables predetermined as of the first period of the planning horizon or lagged random shocks occuring within the planning interval appear as determinants of life cycle resource allocations.

[10] There are three reasons why prices and personal characteristics of the couple might be correlated with fecundity. First, if there are omitted inputs in the reproduction function, the determinants of those inputs will appear to be correlated with the conception rate. Second, even if all endogenous inputs are accounted for, there may exist latent, couple-specific health factors that jointly influence fecundity and the couple's earnings potential or ability to obtain schooling. Third, local-area input prices may be related to the couple's endowments if the couple's choice of location is influenced by such endowments.

including local prices, labor market characteristics, and measures of the availability of public health and family planning services.

The dependent variable measuring fertility is the number of conceptions occurring between the interview dates divided by the months of exposure to the risk of conception, namely, the months in which the wife was not pregnant and/or in which the couple was not abstaining from intercourse.[11] Four fertility control variables are constructed based on the monthly calendar information: the proportions of the total exposure period between 1970 and 1975 (during which the woman was subject to the risk of conception) that the couple was 1) sterilized, 2) using the pill or IUD, 3) using the diaphragm or condom, 4) using "ineffective" techniques (foam, jelly, rhythm, etc.). The grouping of contraceptive methods is based on standard conventions and beliefs on the relative effectiveness of such methods in the U.S. population (Barbara Vaughan et al. 1977; Westoff and Ryder; Bongaarts and Potter).[12] Note, however, that none of these studies has taken into account biases associated with self-selection of birth control techniques.

Also included in the reproduction function, and treated as endogenous are the number of children born by 1970, the monthly frequency of intercourse, and the wife's smoking.[13] The monthly frequency of intercourse is reported in 1970 and 1975, and the average is assumed to apply uniformly throughout the period of exposure. Months of abstinence from intercourse during the five-year period, as noted, are excluded from the period at risk. Smoking by the mother in 1970 may influence reproductive capacity, since it appears that smoking reduces gestation and birth weight, and elevates infant mortality and fetal wastage (see our 1983 article and Kate Praeger et al., 1984), and thus could influence pregnancy wastage and the likelihood of conception. Unfortunately, the amount of smoking during the period is not known, only whether the woman smoked in 1970. (Note in Table 1 that coding is inverted for this variable.)

The set of instruments characterizing the couple's tastes, opportunities, and constraints is extensive; information is employed on the personal characteristics of the couple that were thought to be exogenous with respect to the fertility decision, such as the husband's earnings (not that of the wife's) and the husband's and wife's education, religious affiliation, and age, and features of the residential area that influence employment opportunities and the relative costs of children and other goods, including medical and family planning infrastructure and services, and local sales taxes and prices. (See Table 1 for full list.)

The final sample with complete reproductive histories and sets of characteristics and who were capable of conceiving at the 1970

[11]Exposure to the risk of conception is also reduced by one month following each birth (postpartum amenorrhea). Estimates were also obtained with the postpartum period assumed to be as long as three months for each live birth, but the results were essentially unchanged, with coefficient standard errors somewhat higher. The period during which either partner was sterilized was not subtracted from months of exposure; sterilization is treated as a contraceptive method that, unlike abstinence, has a failure rate.

[12]F-tests were performed to determine if these groups appeared to combine inputs of different effectiveness, but the statistical tests based on the linear, two-stage estimation methodology described below did not support further disaggregation; for example, the effect of no method could not be distinguished from that of unknown or uncertain methods.

[13]The age of the mother, and her age squared, are also included in the reproduction function. Inclusion of the husband's age did not produce a statistically significant effect in the basic specification of the fertility production function. It has been suggested by Barbara Anderson (1975) that in Ireland, prior to 1911, census data indicate that increasing male age is associated with decreasing marital fertility, holding constant for the age of the female. This hypothesis is also employed by Ansley Coale, Anderson, and E. Harm (1979) to account for unexpected increases in age-specific marital fertility rates in central Asian regions of the USSR where the age gap between spouses may have decreased in the twentieth century. We also included the number of children breast fed during the 5-year period as an endogenous determinant of the risk of conception (A. K. Jain and Bongaarts, 1981). However, in none of the two-stage specifications did this variable display any significant effect on the conception rate and it is dropped from the reported analysis.

TABLE 1—VARIABLE DEFINITIONS AND SAMPLE CHARACTERISTICS

Variables	Definition	Mean	Standard Deviation
Endogenous Variables			
Average Monthly Conception Rate, 1970–75	Ratio of the number of conceptions to the number of months of pregnancy risk (exposure period)[a]	.0128	.0181
Sterilization	Proportion of exposure period, 1970–75, that is protected by sterilization	.155	.277
Pill or IUD	Proportion of exposure period, 1970–75, using the pill or IUD.	.408	.397
Diaphragm or Condom	Proportion of exposure period, 1970–75, using the diaphragm or condom	.146	.304
Ineffective Methods	Proportion of exposure period, 1970–75, using jelly, foam, douche, withdrawal, rhythm, or other methods	.132	.281
No Method	Proportion of exposure period, 1970–75, using no method (or do not know)	.160	.286
Births	Number of live births born to mother before 1970	1.95	1.49
Not Smoking	Mother is smoking in 1970 (1 = Yes, 2 = No)	1.65	.477
Coital Frequency	Average monthly coital frequency, 1970 and 1975	8.65	4.91
Exogenous Individual Characteristics			
Education	Years of schooling completed by wife	12.7	2.00
Income 1970	Husband's earned yearly income before taxes in 1970 $	9420	3820
Income 1975	Husband's earned yearly income before taxes in 1975 $	16100	8040
Husband's Age	Age of husband in months in 1970	359.	77.2
Wife's Age	Age of wife in months in 1970	329	69.4
Wife Protestant		.609	.238
Husband Protestant		.600	.241
Wife Catholic		.290	.206
Husband Catholic		.276	.200
Wife Jewish		.0164	.0161
Husband Jewish		.0169	.0166
Wife Mormon		.0327	.0316
Husband Mormon		.0304	.0295
Exogenous Area Characteristics[b]			
Health Expenditures	Local government health and hospital expenditures in thousands of dollars per capita, 1965, at state or SMSA level ($\times 10^2$)	.294	.119
Family Planning in Health Department	Number of health departments with family planning services per capita, 1969, at state or SMSA level ($\times 10^5$)	.907	3.73
Family Planning in Hospitals	Number of hospitals with departments with family planning services per capita, 1969, at state level ($\times 10^5$)	.286	.462
Population per M.D.	Number of persons per medical doctor, 1969, at state or SMSA level	1630	1030
Metropolitan	One if located in SMSA, 0 otherwise	.538	.499
City Size	Population in SMSA in 1970 ($\times 10^{-3}$)	1280	2460
Hospital Beds	Number of hospital beds per capita ($\times 10^2$), 1965, state level	.480	.111
Obstetrician-Gynecologists per Capita	Number of obstetricians-gynecologists per capita, 1969, at state or SMSA level ($\times 10^4$)	.755	.687
Female Unemployment Proportion	Proportion of women in labor force, age 15–59, unemployed, 1970, at state level	.0510	.00978
General Unemployment Proportion	Proportion of the labor force unemployed, 1970, at state level	.0407	.0110
Share of Jobs in Services	Percent of persons employed in services, 1970, at state level ($\times 10$)	75.5	18.0
Share of Jobs Sales	Percent of persons employed in sales, 1970, at state level ($\times 10$)	248	52.8
Share of Jobs in Government	Percent of persons employed in government, 1970, at state level ($\times 10$)	167	61.8
Cigarette Price	Price of cigarettes before taxes, cents per pack, 1967–69, at state level	34.4	3.18
Sales Tax	Retail sales tax on cigarettes. cents per pack, 1967–69, at state level	1.22	.873
Milk Price	Retail price of milk per quart, 1970, at state level	26.9	2.40

[a]The period of exposure approximates the number of months between the two interviews that the woman is exposed to the risk of conception. Those months are excluded during which the couple abstain from intercourse and in which the woman is pregnant. The average period of exposure for the sample is 62 months, and if abstinent months were included, the average period of exposure would be increased by 1.3 percent.

[b]Sources of the area characteristics are described in our 1983 article. Where both state and SMSA characteristics are used, those individuals residing in an SMSA are attributed the characteristics for that SMSA, and those residing outside of a SMSA are attributed the average characteristic for their state.

interview date contains 1753 couples.[14] The average age of the wife was 27.4 years in 1970, and ranges form 15 to 42; the couples had on average 1.9 children in 1970 and had 2.5 children by the second interview in 1975. Descriptive statistics for the working sample and variable definitions are reported in Table 1.

Because the exposure period is relatively short—a maximum of five years—and the overall demand for children among U.S. couples is low, a large proportion of the sample had no conceptions (53 percent). As a consequence, use of least squares, even with instruments, will yield inconsistent estimates of the reproduction technology and thus of the exogenous supply of births among the sample couples.[15] Accordingly, we employ maximum likelihood methods to estimate a limited dependent variable (Tobit) model Because the properties of the Tobit model estimates may be sensitive to the imposed and unverifiable assumption of normality of the underlying distribution of unobservables, the Tobit estimates could be more asymptotically biased than those estimates obtained using the linear specification (Forrest Nelson, 1981). We report coefficient estimates from the linear and censored normal models. Both are estimated with and without instruments to assess the role of selection bias associated with fertility supply variation.

IV. Estimates of the Reproduction Technology

Table 2 reports the ordinary least squares (*OLS*), two-stage least squares (*2SLS*), and single and two-stage Tobit estimates of the reproduction function, with the expected value derivatives derived from the Tobit linear coefficients reported in brackets adjacent to the relevant Tobit index coefficients.[16] The specified nine inputs directly explain (*OLS*) 26 percent of the variation in the monthly conception rate. However, De-Min Wu (1973) and Hausman (1978) specification tests confirm, at the 1 percent level of confidence, that the seven conception inputs, excluding age, are jointly correlated with the "fecundity" residual. Indeed, as predicted, the estimates of the reproductive effects of all of the contraceptive methods are understated by the single-stage estimates compared with the consistent, but somewhat less precise, two-stage estimates designed to deal with contraception selectivity. The estimated effectiveness of the pill/IUD is understated by 100 percent (linear) and 38 percent (Tobit), respectively; the estimated effectiveness of the diaphragm and condom increases by 107 percent in either specification when instruments are used. The most significant bias appears to be imparted to the sterilization coefficient; the two-stage instrumented coefficients are almost 9 times (linear) and over 16 times (Tobit) the size of their single-stage counterparts. Notable is the implication of the two-stage estimates that the diaphragm and condom methods are on a par with the pill and IUD in contraceptive effectiveness as used by a representative couple; both methods, however, are less than half as "use-effective" as sterilization.

The consistent two-stage Tobit point estimates suggest that the sample mean conception rate when no method of contraception is practiced is about 2.5 percent per month.[17] The change in conception rates as-

[14] Excluded were 191 couples (10 percent of observations with complete responses on other characteristics), because either the husband or the wife had already become sterilized prior to the interview in 1970. Exclusion of these couples may introduce some selection bias into our estimates and may lead to an underestimate of the adjustments to and consequences of fertility supply variability.

[15] In particular, the conception probability should be constrained to be nonnegtive for all possible values of the conditioning variables, which is not the case in the linear model.

[16] Estimates of the more flexible Diewert generalization of the linear (Leontief) specification of the reproduction technology were also obtained. Though the values of fertility implied by the two-stage estimates are plausible, the indirect estimation of the collinear inputs does not provide any precise estimates (i.e., low t's) and there is no basis on statistical grounds for preferring this flexible functional form to the functional form reported here.

[17] Underreporting of pregnancy wastage and perhaps induced abortion are a source of measurement error that may render this estimate somewhat lower than the true

TABLE 2—LINEAR AND CENSORED NORMAL (TOBIT) SPECIFICATIONS OF THE REPRODUCTION FUNCTION: SINGLE-STAGE AND TWO-STAGE ESTIMATES

	Linear Model Coefficients		*ML* Tobit Index Coefficients	
Explanatory Variables	*OLS* (1)	*2SLS* (2)	Single Stage (3)	Two Stage (4)
Proportion of Exposure Period Protected by				
Sterilization[a]	−.00453	.0399	−.00448	−.0722
	(2.44)[b]	(2.93)[b]	(1.22)[c]	(3.18)[c]
			[−.00236][d]	[−.0380][d]
Pill or IUD[a]	−.0112	−.0232	−.0229	−.0315
	(7.39)	(2.27)	(7.52)	(1.86)
			[−.0120]	[−.0165]
Diaphragm or Condom[a]	−.00509	−.0161	−.0115	−.0295
	(2.99)	(1.38)	(3.27)	(1.51)
			[−.00604]	[−.0155]
Ineffective Methods[a]	−.00277	−.00861	−.00388	−.00751
	(1.52)	(0.58)	(1.07)	(0.31)
			[−.00204]	[−.00395]
Births before 1970 ($\times 10^2$)[a]	−.0940	.0539	−.184	.0225
	(2.91)	(0.35)	(2.76)	(0.09)
Coital Frequency ($\times 10^{-3}$)[a]	−.143	.122	−.358	.138
	(1.85)	(0.17)	(2.41)	(0.11)
Not Smoking[a]	.00133	.0172	.00266	.0291
	(1.70)	(3.41)	(1.71)	(3.51)
Age of Woman ($\times 10^{-3}$)	−.364	−.284	−.225	−.0481
	(7.14)	(3.01)	(2.09)	(0.30)
Age Squared ($\times 10^{-6}$)	.356	.241	−.0869	−.304
	(4.88)	(1.95)	(0.54)	(1.40)
Intercept	.0994	.0677	−.0967	.0283
	(11.46)	(4.23)	(5.49)	(1.05)
Standard Error	–	–	.0262	.0270
	–	–	(38.99)	(38.67)
R^2	.264	–	–	–
F/ln likelihood	69.45	42.47	1361	1322

[a] Endogenous variable. For instrument list, see Table 1.
[b] Absolute values of t-ratios are shown in parentheses beneath regression coefficients.
[c] Absolute values of asymptotic t-ratios in parentheses beneath regression coefficients.
[d] Bracketed terms are the derivatives of the expected value of the conception rate evaluated at the sample mean values.

sociated with the use of a specific contraceptive based on the estimated Tobit index coefficients suggests that switching from no method to sterilization, the pill or IUD and the diaphragm or condom reduces the probability of conception, given sample usage, to zero, by 55 and 52 percent, respectively. The coefficient for ineffective methods is not statistically different from zero, indicating that such methods are classified appropriately. The two-stage Tobit coefficients for the wife's age are also estimated too imprecisely to depict accurately the age trajectory of potential fertility supply. However, the age variables are jointly statistically significant (χ^2 test, .05 significance level).

Of the other reproduction variables, the number of births accumulated prior to 1970 is negatively correlated with the monthly probability of conception from 1970 to 1975 (*OLS* and single-stage Tobit estimates); however, this pattern is apparently not rooted

"natural" conception rate, and could explain the unexpectedly high effectiveness of the presumptively "less effective" diaphragm/condom methods; in other words, when couples who use these methods fail, they may resort to (unreported) abortions.

in reproductive biology, as neither the two-stage linear nor the Tobit estimates reveal any statistically significant relationship between prior fertility and subsequent conception rates, given contraception. Not smoking appears to increase the conception rate according to the two-stage estimates, as is consistent with recent findings in the medical literature (Praeger et al.). The data do not confirm any significant effect of coital frequency on fertility: the direct association (single stage) is weak but inverse, whereas the two-stage estimates are essentially zero, though of the anticipated sign. The imprecise estimate of the coital frequency effect may be in part due to the inability of the instruments to predict well the variation in this variable ($R^2 = .067$).

V. Fecundity, Fertility, and Life Cycle Adjustments in Contraception, Female Labor Supply, and Earnings

The two-stage Tobit estimates of the effects of contraception on the monthly probability of conception, while interesting by themselves, enable the separation of the behavioral and biological components of fertility so that the consequences of variation in fertility supply can be assessed. The two-stage estimates provide a consistent prediction for each couple of its fertility (conception rate) based on its actual choice of contraceptives. The difference between this consistent prediction, based on the reproduction technology and actual behavior, and the couple's actual conception rate contains the persistent and random components of fertility that are beyond the couple's control, namely, unexplained deviations in fertility supply. These prediction "errors" can be computed for different segments of the life cycle in order to decompose the supply errors into their persistent and random parts.[18]

Because of the truncation at zero of the conception probability, and hence the use of the Tobit estimation procedure, the predicted or expected value of the conception rate n_{ij}^e for each couple j in period i is computed from the formula:

$$n_{ij}^e = E(n_{ij}) = \Phi_{ij} B' F_{ij} + \sigma \phi_{ij}, \tag{15}$$

where B' is the vector of estimated (two-stage) Tobit index coefficients (Table 2, col. 4), F_{ij} is the vector of actual input values for couple j in period i, Φ and ϕ are the cumulative normal and normal density values evaluated on the basis of the couples' actual input values, and σ is the estimated Tobit standard error. A consistent (i.e., as t goes to infinity) estimate of the persistent or fixed component of fertility supply for a couple j for whom fertility, net of inputs, is computed for each of t periods, from (15), is

$$\mu_j = \sum_{i=1}^{t} \left(n_{ij} - n_{ij}^e\right)/t \tag{16}$$

and thus the period-specific, unanticipated fertility component is

$$\varepsilon_{ij} = n_{ij} - n_{ij}^e - \mu_j. \tag{17}$$

To estimate a couple's response to variation in the persistent and transitory supply components, we divide the five-year period containing the calendar information from which the reproduction technology was estimated into two equal two and one-half year segments to compute μ and ε. Our estimates imply that somewhat more than half of the unexplained variation in fertility is persistent; that is, $\sigma_\mu^2/(\sigma_\mu^2 + \sigma_\varepsilon^2) = .56$. While a greater number of interval segments might be desirable for estimating the permanent and transitory components of fertility supply, short intervals provide little information about fecundity in a setting where the average level of contraceptive effectiveness is high, as in our sample.

[18] There are several sources of unexplained variation in birth rates that might be subsumed in μ and the ε_i in equation (1). Statistical noise that arises from errors in measuring the explanatory inputs would be captured in the residuals, along with misspecifications in functional form and omissions of minor inputs that are not correlated with the included inputs. As a consequence, our estimates of the effects of fecundity variation will be biased toward zero.

TABLE 3—EFFECTS OF EXOGENOUS FERTILITY SUPPLY ON ACTUAL FERTILITY IN 1975 AND 1970[a]

	Children Ever Born		
	1975		1970
Explanatory Variables	(1)	(2)	(3)
μ-Persistant Individual Fixed Effect (fecundity)	–	−49.0 (4.37)	−42.6 (5.03)
$\mu \times$ Wife's Age	–	.202 (6.45)	.147 (5.07)
Education	−.104 (5.02)	−.111 (5.72)	−.131 (7.27)
Income 1970 ($\times 10^{-4}$)	.144 (1.26)	.140 (1.30)	.207 (2.07)
Income 1975 ($\times 10^{-4}$)	.0526 (1.02)	.0272 (9.56)	.0150 (0.33)
Wife's Age	.0235 (4.50)	.0261 (5.31)	.0371 (9.66)
Wife's Age Squared ($\times 10^{-4}$)	−.245 (3.90)	−.263 (4.47)	−.407 (7.44)
Husband Protestant	.206 (1.42)	.139 (1.02)	.0345 (0.27)
Husband Catholic	.409 (2.54)	.389 (2.58)	.289 (2.06)
Husband Jewish	−.0650 (0.10)	−.0361 (0.06)	−.00824 (0.01)
Husband Mormon	.775 (1.73)	.619 (1.47)	.281 (0.71)
Female Unemployment Rate	1.61 (0.16)	2.85 (0.31)	−4.02 (0.47)
Family Planning in Health Department	−1839 (1.40)	−2355 (1.92)	−1515 (1.33)
Health Expenditures	99.1 (2.28)	127 (3.11)	100 (2.64)
Intercept	−6.75 (1.99)	−4.93 (1.54)	−4.94 (1.71)
R^2	.181	.282	.412
F	11.85	19.80	35.34

[a]All instruments in Table 1 are included, but only selected socioeconomic variables are reported. Absolute values of t-ratios are shown in parentheses beneath regression coefficients.

If the measure of fecundity μ based on the average of the two intervals from 1970–75 reflects that component of fertility supply that is persistent, then, given imperfect and costly fertility control, variation in μ should explain a significant proportion of the variance in the number of total births born to couples.[19] Moreover, since μ is denominated

[19]The most important properties of the μ and ε estimates are their independence from preferences and their correlation with exogenous variation in the supply of births. Two tests of these propositions were performed. First, we regressed intended births reported in the 1970 interview on the estimated measures of fecundity obtained from the 1970–75 period and on actual fertility in that interval. While this subjective measure of future fertility demand was a statistically significant predictor of actual births occuring in the subsequent 5-year period, it was not statistically correlated with either μ or ε. In the 1975 interview, respondents reported, for each pregnancy, whether or not they had wanted additional children prior to the pregnancy, and if so, whether they might have preferred to have the conception occur later. Regressions of the number of "unwanted" pregnancies and/or the number of errors in the timing of pregnancies on μ and μ interacted with the mother's age indicated that these direct, subjective measures of excess fertility supply were significantly and positively correlated with our indirect (but nonsubjective) measure of couple-specific fecundity.

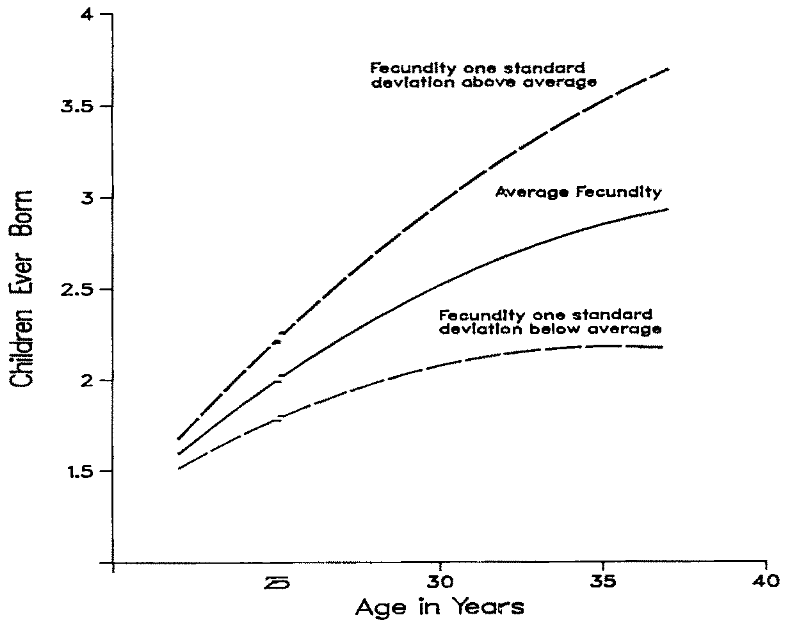


FIGURE 1. CHILDREN EVER BORN BY AGE AND FECUNDITY

as a conception rate, its effect on cumulated births should increase over the life cycle of the couple. Columns 1 and 2 of Table 3 report regressions of the number of children ever born on a set of socioeconomic variables reflecting the demand for births excluding and including, respectively, the constructed measure of couple-specific fecundity and its interaction with the wife's age. As expected, couples with a higher biological propensity to conceive, net of their efforts to reduce conceptions or births, do accumulate increasingly more births as they age, after age 20. Indeed, explanatory power increases by 60 percent when the fertility supply and supply-age interaction variables are added to the set of demand (instrumental) variables. Intercouple heterogeneity in the exogenous supply of births accounts for slightly more than 10 percent of the total variation in cumulated births across families.

Figure 1 plots the cumulative effects on children ever born of variation in μ of one standard deviation around the sample mean, based on the μ coefficients in column 2 of Table 3. As shown, couples with a fecundity level that is one standard deviation above the mean accumulate 0.44 "extra" births by the time the wife is age 30, 0.68 additional births at age 35, and 0.90 additional births when the wife reaches age 40.

The positive association between μ and cumulative fertility that increases with age could be merely an artifact of μ being measured in the last five years of each couple's life cycle. The same regression is reported, therefore, employing children ever born in 1970 as the dependent variable. Fecundity, as measured in the 1970–75 period, should be related to fertility before 1970, if μ represents a persistent biological predisposition to conceive, and column 3 of Table 3 displays essentially the same patterns as when the later measure of fertility is employed. Couples observed to have higher fecundity in the 1970–75 period also appear to have accumulated a greater number of births by 1970, if the wife was then over age 24. We feel justified, therefore, in tentatively interpreting our estimate of μ as the individually persistent component of fertility supply, as a measure of couple fecundity.

A. *Contraceptive Selection*

The shortfall or excess in births associated with fecundity variation is the net result of both the inherent propensities to conceive

TABLE 4—EFFECTS OF EXOGENOUS FERTILITY SUPPLY IN 1970–1972 ON CONTRACEPTIVE USE IN 1973–75, BY METHOD, AND ON FERTILITY DEMAND IN 1975[a]

	Proportion of Period June 1972 to 1975 Using Method:					Demand for Additional Births in 1975
	Sterilization	Pill or IUD	Diaphragm cr Condom	Ineffective Methods	No Method	
Explanatory Variable	*ML* Tobit	*ML* Tobit	*ML* Tobit	*ML* Tobit	*ML* Tobit	*ML* Tobit
μ-Persistant Individual Fixed Effect (fecundity)	14.5	−1.15	5.25	−5.58	.152	−13.0
	(5.52)	(1.06)	(3.02)	(3.25)	(0.14)	(3.07)
ε-Random Effect	−2.64	1.83	−2.49	−3.26	−10.8	1.20
	(1.43)	(1.49)	(1.26)	(1.69)	(7.69)	(0.34)
Intercept	−.445	.101	−.751	−.795	−.418	−2.30
	(2.97)	(4.74)	(14.0)	(14.94)	(9.92)	(13.7)
Standard Error	.970	.745	.996	.958	.714	2.25
	(5.38)	(41.4)	(23.1)	(22.2)	(17.1)	(18.5)
ln likelihood	1364	1651	1093	1000	1095	1040

[a]Absolute values of asymptotic t-ratios are shown in parentheses beneath index coefficients.

and a couple's allocative response to such supply constraints. As the model suggested, whether or not couples can distinguish persistent from random supply shocks, they will react to them over time by reallocating reproductive inputs in order to better equilibrate their realized supply of births to their demand for births. Table 4 reports maximum likelihood Tobit coefficients describing how variations in fecundity (μ) and how random fertility shocks (ε), as measured in the first 2.5 years of the 1970–75 interval, affect 1) the mix of contraceptive methods chosen in the second 2.5-year period, and 2) the demand for additional births at the end of the interval in 1975, as measured by the couples' responses to a question concerning their "intended" number of children. All coefficients are estimated using maximum likelihood Tobit to take account of the concentration of observations at zero for each dependent variable.

The findings in Table 4 indicate that couples with above-average fecundity become sterilized earlier, use the diaphragm or condom more frequently, and use ineffective methods less frequently in the later interval. For given fecundity, couples experiencing an unexpectedly higher conception rate in the first interval spend a smaller proportion of the next interval without any protection and employ ineffective methods less frequently compared to couples who had experienced no fertility "shock." Such couples are also marginally more likely to use the pill or IUD and marginally less likely to become sterilized. The evident reluctance of couples with only a temporary shock to fertility to employ the more permanent and irreversible form of contraception compared to couples with higher fecundity is further confirmation that μ is measuring the more persistent component of fertility supply. Finally, the last column of Table 4 indicates that, as expected, more-fecund couples reduced their demand for additional births or, conversely, subfecund couples had a higher unmet demand for children by 1975.[20]

A natural way to summarize quantitatively the contraceptive adjustments made over the life cycle by couples experiencing different fertility supplies is to weight, for each couple,

[20]While couples make substantial adjustments in their selection of fertility control methods in response to variations in the supply of births, fecundity has no spillover effect on sexual activity among married couples. Neither the frequency of sexual intercourse reported in 1970, nor that reported in 1975, nor the incidence of abstinence were significantly correlated with our measures of exogenous fertility supply or the fertility demand variables. These results are supportive of one of the basic assumptions in Malthus (1798), that there is a "constant passion between the sexes."

TABLE 5—EFFECTS OF EXOGENOUS FERTILITY SUPPLY AND ACTUAL FERTILITY, BY 1972, ON CONTRACEPTIVE EFFECTIVENESS IN 1973–75[a]

	Specification			
Explanatory Variable	(1)	(2)	(3)	(4)
μ-Persistent Individual Fixed Effect (Fecundity)	.137	−.443	–	–
	(10.16)	(5.09)		
$\mu\times$Wife's Age ($\times 10^{-2}$)	–	.180	–	–
		(6.75)		
ε-Random Effect	.0350	.0151	–	–
	(2.28)	(0.15)		
$\varepsilon\times$Wife's Age ($\times 10^{-4}$)	–	.659	–	–
		(0.21)		
Children Ever Born (*CEB*), 1972 ($\times 10^{-2}$)	–	–	.236	.567
			(11.5)	(4.72)
CEB $\times$Wife's Age ($\times 10^{-5}$)	–	–	–	−.814
				(2.78)
Children Born (*CB*), 1970–72 ($\times 10^{-3}$)	–	–	.373	−.927
			(0.72)	(0.29)
CB $\times$Wife's Age ($\times 10^{-5}$)	–	–	–	.205
				(0.22)
Education ($\times 10^{-3}$)	−.434	−.487	−.143	−.0771
	(2.78)	(3.16)	(0.91)	(0.49)
Income 1970 ($\times 10^{-7}$)	.102	.0941	.490	.310
	(1.19)	(1.10)	(0.57)	(0.36)
Income 1975 ($\times 10^{-7}$)	.0013	.0493	.153	.147
	(0.29)	(0.13)	(0.40)	(0.38)
Wife's Age ($\times 10^{-3}$)	.204	.198	.112	.0710
	(5.60)	(5.50)	(3.01)	(1.50)
Wife's Age Squared ($\times 10^{-6}$)	−.239	−.230	−.143	−.0696
	(5.05)	(4.91)	(2.99)	(1.09)
Husband Protestant ($\times 10^{-2}$)	−.0917	−.0885	−.0792	−.0734
	(0.84)	(0.82)	(0.73)	(0.68)
Husband Catholic ($\times 10^{-2}$)	−.0602	−.0749	−1.42	−1.31
	(0.50)	(0.63)	(1.19)	(1.10)
Husband Jewish ($\times 10^{-2}$)	−.917	−8.78	−7.88	−7.65
	(1.96)	(1.91)	(1.70)	(1.66)
Husband Mormon ($\times 10^{2}$)	−.511	−4.35	−5.12	−5.14
	(1.51)	(1.31)	(1.53)	(1.54)
Female Unemployment Rate	−.0219	−.0235	−.0357	−.0365
	(0.30)	(0.32)	(0.49)	(0.50)
Family Planning in Heath Department	150	146	99.1	92.9
	(2.51)	(2.47)	(1.68)	(1.59)
Health Expenditures	−.148	−.0737	−.512	−.546
	(0.45)	(0.23)	(1.58)	(1.69)
Intercept	.0170	.0154	.0202	.0250
	(0.67)	(0.62)	(0.81)	(0.99)
R^2	.118	.145	.138	.143
F	6.77	8.10	8.11	7.95

[a]All instruments in Table 1 are included, but only selected socioeconomic variables are reported. Absolute values of t-ratios are shown in parentheses beneath regression coefficients.

the contraceptive effectiveness of each of the methods, as estimated in column 4, Table 2, by the proportions of the interval each method was employed by the couple. Table 5 reports regressions of this measure of the average effectiveness of contraceptive practices employed in the 1973–75 interval on μ and ε, as measured in the 1970–72 interval, and on the set of socioeconomic variables. In column 1, Table 5, the hypotheses are con-

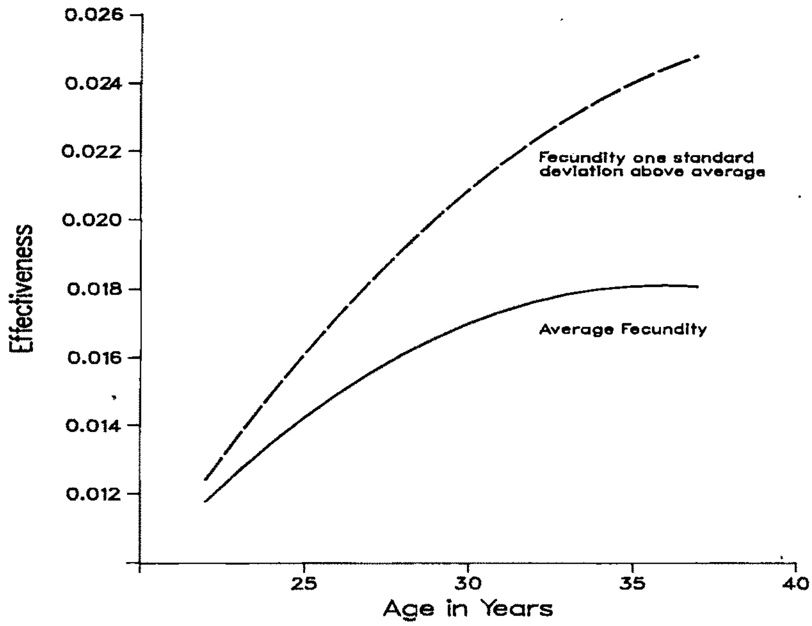


FIGURE 2. CONTRACEPTIVE EFFECTIVENESS BY AGE AND FECUNDITY

firmed at the 1 percent level that (*i*) couples with higher fecundity choose more efficient methods and (*ii*) the reallocation of methods is stronger when μ varies than when there are changes in the serially uncorrelated supply of births, ε. In column 2, the hypothesis that the adjustments to persistent fertility supply effects grow stronger over the life cycle is also confirmed—as higher-μ couples age, they adjust upward their contraceptive effectiveness.[21] The point estimates indicate, as plotted in Figure 2, that couples whose fecundity is one standard deviation higher than the sample mean exhibit an average contraceptive effectiveness that is 13 percent higher when the wife is age 25, 23 percent higher when she is 30, and 45 percent higher when she is 40, compared to couples of the same age with average fecundity. As shown in Table 3 and Figure 1, however, these adjustments are not sufficient to offset fully the persistently higher conception rate. The offset to the higher supply level is 10 percent by age 25 and reaches only 43 percent by age 40. Unwanted births according to these estimates are still a substantial share of all U.S. births in 1970–75, as is consistent with subjectively ascertained responses from fertility surveys (Westoff and Ryder).

In columns 3 and 4 of Table 5, the persistant and temporary measures of fertility supply are replaced by the couple's actual fertility—its cumulated births through 1972 and actual births occuring in the interval 1970–72—in order to assess how well actual prior fertility serves as a proxy for exogenous fertility supply. This comparison enables us to measure the bias from, or the conse-

[21] Duration of marriage might provide a more accurate indicator of the couple's accumulated knowledge of their fecundity than would the wife's age. However, there is no reason to believe that age at marriage is unrelated to preferences for family size and/or for life cycle labor supply. Testing this hypothesis, anyway, we used duration of marriage and duration of marriage interacted with our measure of fecundity in the contraceptive choice equations, analogous to the specifications reported in Table 5. In all specifications, the coefficients on "marriage duration" were statistically significant at the .1 percent level. The explanatory power of the regression, corrected for degrees of freedom, exceeded that reported in the age-fecundity interaction specification reported in Table 5, with the positive effect of fecundity on contraceptive efficiency increasing after 19 months of marriage.

quences of, specifying actual fertility as a variable conditioning household decisions, when fertility itself also embodies the effects of household preferences. As discussed in Section II, the existence of a serial correlation in couple-specific preferences should bias downward the effect of fertility supply variation on contraceptive choice when actual fertility is used to represent exogenous supply. This is confirmed in Table 5. The estimates of Table 3 indicate that μ must be higher by .0427 in order for cumulated births to increase by one for a woman whose age is at the sample mean in 1972 (age 30). The estimates of column 2 in Table 5 indicate that this increase in fecundity, sufficient to raise actual fertility by one birth, would increase contraceptive effectiveness by .0087 or by 54 percent. In contrast, a one-birth rise in actual fertility (column 4) is associated with only a .0028 or 17 percent increase in contraceptive effectiveness at the sample means. The life cycle adjustment in contraceptive choice to fertility supply variation, as measured by actual fertility, is less than one-third that measured by using μ, which presumably is purged of any fertility-preference component. The magnitude of bias is, therefore, considerable in treating fertility as exogenous in the study of contraceptive practice.

The understatement of the dynamic adjustments in contraception to fecundity variation is not the only consequence of regressing, essentially, current fertility demand variables (i.e., contraception) on lagged fertility, as in most previous studies of contraceptive choice (Michael; Michael and Willis). Since actual fertility, because it is partially demand-determined, is conditioned by prices and characteristics of the parents, lagged fertility will pick up the effects of these fertility determinants on current contraceptive choice. This is also seen in Table 5. In column 2, in which μ and ε are included, the two most significant "demand" determinants of the effectiveness of contraception, aside from age, are the wife's education and the extent of local family planning programs.[22] In the specifications in which measures of prior fertility are included as regressors, however, the coefficients on these two variables are reduced substantially: the wife's schooling coefficient is reduced in absolute value by 75 percent and loses its statistical significance; the statistically significant positive family planning coefficient is reduced in absolute value by a third.

B. *Female Life Cycle Labor Supply and Earnings*

In the second half of the five-year interval between the 1970 and 1975 interview dates, 64 percent of the sample wives worked at least four days per week in one month. Wives' participation in the labor market, according to this definition, rose from 51 percent in any month during 1973 to 54 percent in the year preceding the last interview in 1975. Columns 1, 3, and 5 in Table 6 report Tobit maximum likelihood estimates of the effects of variations in the persistent and transitory measures of fertility supply on these measures of the wife's labor supply.[23] In all equations, the fertility supply variables add significantly to the likelihood function. The estimates indicate that among couples whose fecundity is one standard deviation above the population mean, the proportion of

[22] Note that in the 1970–75 period, more-educated mothers use effective techniques less often; however, they also have distinctly fewer births before 1970 (Table 3, col. 3), and thus appear to be headed for a smaller completed family size. It is widely observed that more-educated women marry later and experience a period of "catching up" with their cumulative fertility, (Rosenzweig and Daniel Seiver, 1982); consistent with this, a majority of the more educated mothers appear to be employing their knowledge of the reproductive process (i.e., technology) to increase their birth rate rather than decrease it in the period observed from 1970 to 1975.

[23] Because of the limitations of the Tobit and Probit computer programs, we could not include all of the exogenous instruments in the labor supply equations. The variables reported in Table 6 thus are the complete set of "regressors." Least squares estimates of the labor supply equations, including and excluding the variables excluded from the Tobit and Probit models, suggest that the results are only trivially affected. Similar tests performed on the children ever born and contraceptive efficiency equations confirmed the robustness of the results to this change in specification.

TABLE 6—EFFECTS OF EXOGENOUS FERTILITY SUPPLY AND ACTUAL FERTILITY ON SUBSEQUENT MEASURES OF FEMALE LABOR SUPPLY[a]

	Proportion of Interval Employed in 1973–1975		Participation in 1973		Participation in 1975	
	ML Tobit		*ML* Probit		*ML* Probit	
Explanatory Variable	(1)	(2)	(3)	(4)	(5)	(6)
μ-Persistant Individual Fixed Effect (fecundity)	−5.55 (6.75)	–	−7.84 (4.48)	–	−12.4 (6.90)	–
ε-Random Effect	.162 (0.17)	–	−4.50 (2.27)	–	4.13 (2.04)	–
Children Ever Born, 1970	–	−.00371 (0.31)	–	−.0411 (1.56)	–	.0369 (1.41)
Children Born 1970–72	–	−.236 (8.28)	–	−.502 (8.14)	–	−.388 (6.42)
Education	.0545 (7.07)	.0555 (7.03)	.115 (6.81)	.116 (6.62)	.107 (6.34)	.113 (6.53)
Income 1970 ($\times 10^{-4}$)	−.0678 (1.35)	−.0638 (1.27)	−.237 (2.19)	−.231 (2.11)	−.149 (1.38)	−.138 (1.27)
Income 1975 ($\times 10^{-4}$)	−.155 (6.68)	−.156 (6.71)	−.238 (4.79)	−.243 (4.85)	−.259 (5.25)	−.261 (5.29)
Wife's Age	−.00232 (1.09)	−.00421 (1.92)	−.0117 (2.54)	−.0136 (2.83)	.00330 (0.72)	−.00197 (0.42)
Wife's Age Squared ($\times 10^{-4}$)	.0443 (1.59)	.0606 (2.14)	.169 (2.80)	.181 (2.92)	.0145 (0.24)	.0338 (0.55)
Husband Protestant	−.126 (2.35)	−.116 (2.19)	−.218 (1.85)	−.204 (1.71)	−.199 (1.67)	−.189 (7.59)
Husband Catholic	−.182 (3.15)	−.165 (2.85)	−.258 (2.03)	−.213 (1.65)	−.298 (2.33)	−.300 (2.35)
Husband Jewish	−.0706 (0.58)	.0930 (0.76)	−.0937 (0.35)	−.169 (0.63)	.0588 (0.22)	.0495 (0.18)
Husband Mormon	−2.19 (2.22)	−.190 (1.94)	−.593 (2.77)	−.510 (2.35)	−.433 (2.06)	−.425 (2.03)
Female Unemployment Rate	−1.02 (0.65)	−.908 (0.59)	−5.32 (1.59)	−5.30 (1.58)	−1.47 (0.44)	−1.19 (0.35)
Family Planning in Health Department	1871. (0.60)	2616. (0.84)	−3641. (0.54)	−1776. (0.26)	2706. (0.40)	3719. (0.55)
Health Expenditures	−3.96 (0.31)	−3.34 (0.26)	13.6 (0.50)	18.3 (0.67)	1.16 (0.04)	.362 (0.13)
Intercept	.320 (0.79)	.850 (2.04)	1.58 (1.80)	2.33 (2.55)	−1.42 (1.63)	−.183 (0.20)
Standard Error	.563 (49.2)	.560 (48.9)	–	–	–	–
ln likelihood	1475	1465	1150	1130	1139	1142

[a]Absolute values of asymptotic t-ratios are shown in parentheses beneath slope coefficients.

months between 1973 and 1975 in which the wife participated in the labor market was reduced by 16 percent. As was evident for fertility and contraception, moreover, the effects of fecundity variation appear to increase over the life cycle—among the highly fecund couples, the probability that the wife participated in the labor market was 12 percent lower in 1973 and was 17 percent lower in 1975.

In contrast to the persistent fecundity effects embodied in μ, a transitory increase in the supply of births appears to have little or no lasting effect on the proportion of time spent by the wife in the labor market in the 2.5-year interval following the unanticipated birth (col. 1). While the wife's probability of participation falls in the year immediately following the interval during which the fertility supply shock occurs (col. 3), the likeli-

hood of her participating increases by almost as much two years later (col. 3). This compensatory pattern of life cycle labor supply response to exogenous, transitory changes in fertility is similar to that found by Rosenzweig and Wolpin in their study of the labor supply responses of women experiencing an (unanticipated) multiple birth in their first pregnancy.

Estimates of the associations between actual cumulative and recent fertility and female labor supply (cols. 2, 4, and 6, Table 6) provide a different picture of the causal relationship between the timing of labor supply and exogenous changes in fertility. The number of children ever born prior to 1970 has no significant association with the wife's labor market time in the 1973–75 period; higher cumulated births prior to 1970 are associated with a marginally lower participation probability in 1973, but are associated with a marginally higher participation probability in 1975. The number of children born between 1970 and 1972, however, is significantly and negatively associated with time worked by the wife in the 1973–75 period and with labor force participation probabilities in both 1973 and 1975

These labor supply-fertility results, based on actual fertility values, are often interpreted as reflecting the greater "time intensity" of younger children (Willis). However, in contrast with μ and ε, which are independent of preferences by construction, actual fertility will be correlated with latent taste factors that also influence labor supply. It is thus not possible to rule out the hypothesis, for example, that women who postpone births do so in order to participate in the labor market while young. With respect to quantitative effects, the associations between actual fertility in the 1970–72 period and the wife's participation in 1973 and 1975 overstate both the immediate (1973) negative and longer-term labor supply responses to an unanticipated birth. The wife's probability of labor force participation is reduced at the sample means by 40 percent in 1973, and by 28 percent in 1975, in response to an *actual* birth between 1970–72. The arrival of an *unanticipated* birth in the same 30-month period (less 9 months for the pregnancy) would require an increase in the monthly conception rate of approximately 1/21; this increase in ε sufficient to yield one unanticipated extra birth would reduce the wife's likelihood of participation by only 18 percent in the following year (1973), and *increases* the participation probability by 14 percent in 1975.[24]

Given heterogeneity in preferences, it would be expected that the use of actual fertility might lead to negatively biased estimates of labor supply responses to exogenous variations in fertility supply. However, the estimates in Table 6 suggest that none of the other estimated coefficients in the labor supply equations are importantly affected by whether or not actual fertility variables are included as regressors in place of the fertility supply variables. This is in marked contrast to the estimates of the contraception equations.

One important labor supply determinant left out of the equations reported in Table 6, and likely to be correlated with fertility, is the wife's wage. This was deliberate, since we regard the wage rate received by the wife as endogenously determined, a function of work experience and thus of prior labor supply decisions. Our estimates thus show the reduced-form effects of exogenous variation in fertility supply on life cycle allocation decisions. We can, however, also examine the net effect of unanticipated fertility variation on the earnings of the wife. If children and the wife's home time are complements, as our labor supply estimates suggest, and there are important payoffs to work experience, then we would expect that more fecund wives would have less labor force experience and thus would receive a lower wage rate at any age. Variation in the supply of births should also account for a significant proportion of the wage variation among married women.

Jacob Mincer and Solomon Polachek estimated the effect of fertility on the earnings of married women using actual cumulative fertility as a regressor. While they

[24] This value is derived as follows: $(1/21)(\beta_\varepsilon\phi(Z'\beta)) = .0972$ for 1973, where β_ε = the estimated ε coefficient in the probit equation and ϕ is the normal density evaluated at the sample means of the variables Z using the vector of estimated probit coefficients β.

TABLE 7—EFFECTS OF EXOGENOUS FERTILITY SUPPLY AND ACTUAL FERTILITY ON (LOG) WEEKLY EARNINGS OF MARRIED WOMEN, 1975[a]

Explanatory Variable:	Specification (1)	(2)	(3)
μ-Persistent Individual Fixed Effect (fecundity)	–	−5.28 (3.20)	–
Children Ever Born, 1975	–	–	−.0885 (4.27)
Education	.0632 (4.22)	.0648 (4.36)	.0533 (3.55)
Income 1970 ($\times 10^{-4}$)	−.128 (2.57)	−.134 (2.72)	−.133 (2.70)
Income 1975 ($\times 10^{-4}$)	−1.56 (1.50)	−.155 (1.50)	−.151 (1.47)
Wife's Age ($\times 10^{-2}$)	−.279 (0.61)	−.409 (0.90)	−.111 (0.24)
Wife's Age Squared ($\times 10^{-5}$)	.596 (1.09)	.753 (.137)	.461 (0.85)
Husband Protestant	−.0481 (0.46)	−.0522 (0.50)	−.0348 (0.34)
Husband Catholic	−.114 (0.99)	−.122 (1.06)	−.0506 (0.44)
Husband Jewish	.137 (0.61)	.146 (0.65)	.148 (0.67)
Husband Mormon	−.383 (1.93)	−.350 (1.77)	−.284 (1.44)
Female Unemployment Rate	5.76 (1.87)	5.69 (1.86)	5.62 (1.84)
Family Planning in Health Department	3200 (0.53)	3808 (0.63)	2851 (0.48)
Health Expenditures	−10.8 (0.44)	−10.4 (0.42)	−5.59 (0.23)
Intercept	3.71 (3.95)	3.95 (4.22)	3.58 (3.84)
R^2	.0518	.0618	.0695
F	4.36	4.86	5.50

[a]All instruments in Table 1 included, but only selected socioeconomic variables are reported. Absolute values of t-ratios are shown in parentheses beneath slope coefficients.

recognized the endogeneity of fertility, simultaneous equations methods were eschewed (correctly) because, as they noted, "...the fertility function would be estimated by the same variables as the labor supply function" (1974, p. S99). Our procedure, which takes account of the reproductive technology, permits the estimation of the effects of exogenous variation in the supply of births on the earnings of married women without the need to impose arbitrary identification restrictions on the household demand system.

Table 7 reports estimates from three specifications of the determinants of the log of the weekly earnings for the subset of 972 wives working in 1975. Since human capital theory indicates that an individual's current wage is a function of his or her labor force experience, all determinants of the wife's labor supply are included in each specification. Note that this means that correction for selectivity bias resulting from self-selection by women into the labor market must rely on differences in functional form between the sample inclusion (labor force participation) and wage equations or on distributional assumptions. Use of the probit procedure to estimate the participation equation allows one test (Heckman, 1974) for selectivity bias.

On the basis of this, we could not reject the hypothesis of no selectivity when either μ or actual fertility was included in the earnings equation (specifications (2) and (3)). The hypothesis was rejected for the specification excluding fertility; however, coefficient estimates were essentially identical whether or not selectivity was "corrected" and the uncorrected estimates are reported for comparability across specifications in Table 7.

The findings reported in Table 7 confirm that inclusion of the fecundity variable increases the explanatory power of the female earnings regression by 20 percent; persistent fertility supply variation alone accounts for about 1 percent of the total variation in earnings among married women. The coefficient on μ implies that a one standard deviation increase in fecundity is associated with a 10 percent loss in the wife's weekly earnings, with the absolute amount of the earnings loss increasing over the life cycle (since earnings rise). Variations in fecundity combined with imperfect and costly fertility control thus exert substantial effects on the distribution of market human capital accumulated by American wives, presumably by varying their experience and training the the labor market.

Actual cumulative fertility and the wife's earnings are also significantly and negatively associated. However, the coefficient on children ever born understates the decrease in female earnings due to variation in fertility supply—a rise in fecundity sufficient to yield an increase in one unanticipated or "excess" birth (from Table 3) for women aged 33 (the mean age of the sample women in 1975) would lower earnings by 17 percent compared to the 9 percent reduction associated with a one-child increase in children ever born. Conditioning earnings on actual fertility also biases downward the coefficient on the wife's schooling in her earnings function by 18 percent.

Both of the estimates of the effects of fertility on the earnings of married women in Table 7 exceed those found by Mincer and Polachek; however, their earnings specifications included as regressors prior work experience. When predicted work experience variables were used by them, in recognition of the endogeneity of labor supply, the coefficient on children ever born became positive. The positive relationship between actual fertility, net of cumulative labor supply effects, and female earnings found by Mincer and Polachek suggests that an (exogenous) increase in female earnings may increase the demand for children; use of actual fertility, reflecting both the demand for and supply of births, thus underpredicts the earnings loss from purely supply-induced or "unwanted" variations in fertility.

VI. Conclusion

In our sample of white married women in the United States in 1970, 27 percent of the couples reported that they had one or more unwanted children by 1975, while 62 percent reported that they had experienced one or more births earlier than they had intended. Thus, in a society that is relatively sophisticated with respect to contraception, the number and timing of births serves to constrain resource allocations for a significant proportion of households, and the stochastic nature of fertility is an important aspect of their decision making. In this paper, we have developed a methodology for disentangling empirically in households the biologically determined supply of births from the demand for births in order to assess the consequences of exogenous variation in fertility supply on household fertility behavior and their household allocation decisions. Our estimates indicate that variation in fertility reflects heterogeneity in both the preferences of households and inherent, couple-specific reproductive potential, with one-third of the explained variance in births due to heterogeneity in supply. Neither studies that have viewed fertility solely as an expression of individual choice, constrained and conditioned by prices and income, nor studies that have treated family size and age-composition as purely exogenous determinants of behavior, thus provide appropriate estimates of the effects of fertility supply on household behavior or even of the consequences of contraception for the birth rate.

Our findings suggest that couples alter their mix of contraceptive methods over their life cycle in response to their fertility supply

experience. As a consequence of this dynamic behavior, estimates of the effectiveness of contraceptive methods that ignore heterogeneity in fertility supply significantly understate their effectiveness in reducing conceptions. Life cycle patterns of labor force participation by married women are also shown to be sensitive to variations in the exogenous supply of births, and married women who experience excess fertility due to the lack of costless or perfect control over fertility supply incur a substantial loss in earnings when they enter the labor force.

Our attempt to isolate that component of fertility that is independent of preferences required estimation of the direct biological/technical relationships between endogenous and exogenous factors influencing the production of births. Our results thus could be sensitive to the specification and estimation of the reproductive technology. The data and approach appear to confirm, however, the importance of fecundity variation in motivating couples to pursue distinctive contraceptive and labor supply strategies. They also suggest the potential importance of the costs of and access to contraceptive technology in accounting for patterns of resource allocations among households, which should be of even greater importance in low-income countries where contraceptive knowledge may be less pervasive.

REFERENCES

Anderson, Barbara A., "Male Age and Fertility, Results from Ireland Prior to 1911," *Population Index*, July 1975, *41*, 561–66.

Bongaarts, John, "Fertility Inhibiting Effects of the Intermediate Fertility Variables," *Studies in Family Planning*, April 1982, *13*, 179–89.

______ **and Menken, Jane**, "The Supply of Children: A Critical Essay," in R. A. Bulatao and R. D. Lee, eds., *Determinants of Fertility in Developing Countries*, Vol. 1, New York: Academic Press, 1983, 27–60.

______ **and Potter, Robert G.**, *Fertility, Biology and Behavior*, New York: Academic Press, 1983.

Boulier, Bryan and Mankiw, N. G., "Econometric Model of the Demand, Supply and Regulation of Fertility," mimeo., Office of Population Research, Princeton University, September 1980.

Bulatao, Rodolfo A. and Lee, Ronald D., *Determinants of Fertility in Developing Countries*, Vol. 2, New York: Academic Press, 1983.

Butz, William P. and Ward, Michael P., "The Emergence of Countercyclical U.S. Fertility," *American Economic Review*, September 1979, *69*, 318–28.

Coale, Ansley J., Anderson, Barbara A. and Harm, E., *Human Fertility in Russia Since the Nineteenth Century*, Princeton: Princeton University Press, 1979.

Deaton, Angus, "Inequality and Needs: Some Experimental Results for Sri Lanka," in Yoram Ben-Porath, ed., *Income Distribution and the Family, Population and Development Review*, September 1982, Suppl., *8*, 35–52.

Easterlin, Richard A., Pollak, Robert A. and Wachter, Michael L., "Toward a More General Economic Model of Fertility Determination," in Richard A. Easterlin, ed., *Population and Economic Change in Developing Countries*, Chicago: University of Chicago Press, 1980.

______ **and Crimmins, Eileen M.**, "An Exploratory Study of the 'Synthesis Framework' of Fertility Determination, with WFS Data," *World Fertility Survey Scientific Report*, No. 40, Voorburg: International Statistical Institute, November 1982.

Fleisher, Belton M. and Rhodes, George F., Jr., "Fertility, Women's Wage Rates, and Labor Supply," *American Economic Review*, March 1979, *69*, 14–24.

Hausman, Jerry A., "Specification Tests in Econometrics," *Econometrica*, November 1978, *46*, 1251–71.

______ **and Ruud, Paul**, "Family Labor Supply with Taxes," *American Economic Review Proceedings*, May 1984, *74*, 242–48.

Heckman, James J., "Sample Bias as a Specification Error," *Econometrica*, January 1974, *47*, 153–62.

______ **and MaCurdy, Thomas**, "A Life Cycle Model of Female Labour Supply," *Review of Economic Studies*, November 1980, *47*, 47–74.

______ and Willis, Robert, "Estimation of a Stochastic Model of Reproduction: An Econometric Approach," in N. E. Terleckyj, ed., *Household Production and Consumption*, NBER *Studies in Income and Wealth*, No. 40, New York: Columbia University Press, 1975, 99–138.

Jain, A. K. and Bongaarts, John, "Breastfeeding: Patterns, Correlates and Fertility Effects," *Studies in Family Planning*, January 1981, *12*, 79–108.

McFadden, Daniel, "Comments on Estimation of a Stochastic Model of Reproduction: An Econometric Approach," in N. E. Terleckyj, Ed., *Household Production and Consumption*, NBER, New York: Columbia University Press, 1975, 139–45.

Malthus, Thomas R., *An Essay on the Principle of Population*, Baltimore: Penguin, 1979 (original 1798).

Michael, Robert T., "Education and the Derived Demand for Children," *Journal of Political Economy*, March/April 1973, *81*, S128–64.

______ and Willis, Robert J., "Contraception and Fertility: Household Production Under Uncertainty," in N. E. Terleckyj, ed., *Household Production and Consumption*, NBER, New York: Columbia University Press, 1975, 27–93.

Mincer, Jacob and Polachek, Solomon, "Family Investments in Human Capital: Earnings of Women," *Journal of Political Economy*, March/April 1974, *82*, S76–108.

Nelson, Forrest D., "A Test for Misspecification in the Censored Normal Model," *Econometrica*, July 1981, *49*, 1317–30.

Nerlove, Marc and Schultz, T. Paul, *Love and Life Between the Censuses: A Model of Family Decision-Making in Puerto Rico*, Santa Monica: Rand Corporation, 1970.

Pollak, Robert A. and Wales, Terence, "Comparison of the Quadratic Expenditure System and Translog Demand Systems with Alternative Specifications of Demographic Effects," *Econometrica*, May 1980, *48*, 595–612.

Praeger, Kate et al., "Smoking and Drinking Behavior Before and During Pregnancy of Married Mothers of Live-born Infants and Still-born Infants," *Public Health Reports*, March/April 1984, *99*, 117–27.

Rosenzweig, Mark R. and Evenson, Robert E., "Fertility, Schooling and the Economic Contribution of Children in Rural India: An Econometric Analysis," *Econometrica*, September 1977, *45*, 1065–80.

______ and Schultz, T. Paul, "Estimating a Household Production Function: Heterogeneity, the Demand for Health Inputs and Their Effects on Birthweight," *Journal of Political Economy*, October 1983, *91*, 723–46.

______ and Seiver, Daniel, "Education and Contraceptive Choice: A Conditional Demand Framework," *International Economic Review*, February 1982, *23*, 171–98.

______ and Wolpin, Kenneth I., "Life Cycle Labor Supply and Fertility," *Journal of Political Economy*, April 1980, *89*, 1059–85.

Schultz, T. Paul, "An Economic Model of Family Planning and Fertility," *Journal of Political Economy*, March 1969, *77*, 153–80.

______, "Estimating Labor Supply Functions for Married Women," in J. P. Smith, ed., *Female Labor Supply*, Princeton: Princeton University Press, 1980.

Smith, James P., *Female Labor Supply*, Princeton: Princeton University Press, 1980.

Vaughn, Barbara et al., "Contraceptive Failure Among Married Women in the U.S.: 1970–73," *Family Planning Perspectives*, June 1977, *9*, 251–58.

Westoff, Charles F. and Ryder, Norman B., *The Contraceptive Revolution*, Princeton: Princeton University Press, 1977.

Willis, Robert J., "A New Approach to the Economic Theory of Fertility Behavior," *Journal of Political Economy*, March/April 1973, *81*, S14–64.

Wolpin, Kenneth I., "An Estimable Dynamic Stochastic Model of Fertility and Child Mortality," *Journal of Political Economy*, October 1984, *92*, 852–74.

Wu, De-Min, "Alternative Tests of Independence Between Stochastic Regressors and Disturbances," *Econometrica*, July 1973, *41*, 733–500.

Earnings, Schooling, Ability, and Cognitive Skills

By M. BOISSIERE, J. B. KNIGHT, AND R. H. SABOT*

Conventional estimates now available for a large number of countries generally indicate that the social returns to education are positive, large, and competitive with returns to investment in physical capital.[1] That such estimates are good guides for public resource allocation has, however, been questioned. The heart of the problem lies in the interpretation of the positive relationship between the education and the earnings of workers: whether, as the conventional estimates assume, the coefficient of the education variable in the earnings function measures the effect on the productivity of workers of human capital acquired in school. It has been hypothesized that education in part, or instead, represents screening for native ability and motivation, or credentialism, and that as a consequence conventional measures of the social benefit of education are substantially upward biased.[2]

In this paper we attempt to distinguish the influence on earnings of cognitive achievement, native ability, and years of education as a means of adjudicating the human capital, screening, and credentialist hypotheses. Our econometric analysis is based on two rigorously comparable micro data sets from Kenya and Tanzania, generated by surveys of the urban wage-labor force specifically for this study. These data sets contain the usual variables found in earnings function estimates of the benefits of schooling—individual earnings, years of education, and years of employment experience. In addition, they contain two variables—measures of the worker's cognitive skills and of his or her reasoning ability—not previously found in studies of developing countries and only rarely found in studies of the education-earnings relationship in developed countries.[3] With these variables we can estimate the direct effects on earnings of cognitive skills, ability, and years of schooling. By using them to estimate educational production functions and educational attainment functions, and linking these functions with the earnings function in a recursive framework, we can also assess the various indirect effects on earnings of ability and years of schooling. Having data sets from two countries very similar with respect to size, resource endowments, structure of production and employment, and level of development means that not only can we subject our results to the usual statistical tests, but we can also assess their replicability.

Both Kenya and Tanzania have nearly achieved the objective of universal primary education while university enrollments remain at less than 1 percent of the relevant age group. The important policy issues re-

*Boissiere: Development Research Department, The World Bank, 1818 H Street, Washington, D.C. 20433; Knight: Institute of Economics and Statistics, Oxford University; Sabot: Williams College, Williamstown, MA 01267. We are grateful to the Educational Testing Service of Princeton for the design of tests used in this study and to J. Armitage, J. Behrman, J. Hausman, D. Hendry, D. Jamison, and an anonymous referee for their insights and advice. Helpful comments were also received from participants in seminars at Oxford and Yale universities. The views presented here are our own; they should not be interpreted as reflecting those of the World Bank.

[1] George Psacharopoulos (1973; 1981) contains a listing of 44 countries in which rate of return studies had been conducted and of the estimates obtained.

[2] For instance, Kenneth Arrow (1973), Mark Blaug (1976), Samuel Bowles and Herbert Gintis (1976), John Riley (1979), Michael Spence (1976), and Lester Thurow (1975).

[3] For attempts to control for ability and/or for cognitive achievement in studies for the United States, see Jere Behrman et al. (1980), Gary Chamberlain and Zvi Griliches (1977), Griliches and William Mason (1972), Michael Olneck (1977), Paul Taubman and Terence Wales (1974), Taubman (1975), and David Wise (1975); see also the survey articles by Griliches (1977; 1979). In most instances the data refer to special subgroups in the population and clear distinction cannot be made between natural ability and cognitive skills acquired in school.

garding mass education in East Africa arise at the secondary level. We are therefore concerned to evaluate the benefits of secondary education. The public educational system in both countries is meritocratic and years of education may thus provide good signals of ability. The public sector is an influential employer of urban labor in East Africa, accounting in 1980 for 39 and 61 percent of the total in Kenya and Tanzania, respectively. Moreover, institutional arrangements suggest that access to public sector employment grades and entry pay are influenced by educational qualifications. Explanations of the earnings-education relationship in terms of screening or credentialism cannot therefore be dismissed in the present context.

Section II presents our recursive model of ability, years of education, cognitive achievement, and earnings. Section III discusses our data. Estimates of the model are presented—earnings functions in Section IV and educational production and attainment functions in Section V—and their implications for the human capital, screening and credentialist hypotheses are discussed.

I. The Model

In the conventional measurement of the social rate of return to (say) secondary education, the benefit stream is estimated by means of an earnings function, of which the following, for a sample of primary and secondary school completers, is an example:

$$(1) \qquad \ln W = a + bS + cL + dL^2 + u$$

where $\ln W$ = log of (pre-tax) earnings of the individual, S = dummy variable signifying that the individual has precisely completed secondary education, individuals with a complete primary education forming the base subcategory,[4] L = the number of years of employment experience of the individual, and u = a disturbance term.

[4]Schooling is introduced as a dichotomous rather than a continuous variable for reasons of survey design, to be discussed below.

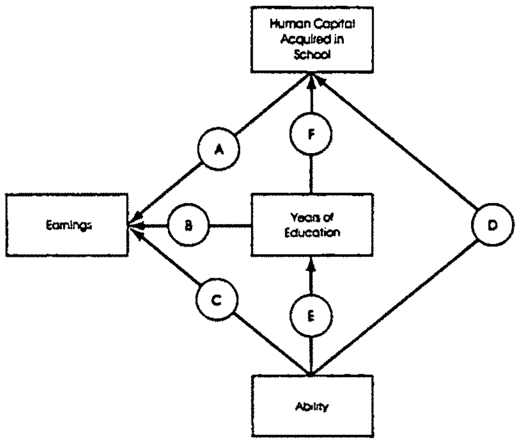


FIGURE 1

The term S is interpreted as a proxy for the cognitive skills or other marketable traits acquired in secondary education. The cross-section earnings function is used to simulate two time-series, $\hat{W}_p$ and $\hat{W}_s$, representing the predicted earnings, over their expected working-lives, of primary and secondary school leavers, respectively. The difference between the educational groups in predicted lifetime earnings is then used as the estimate of the social benefits of secondary education, from which the rate of return can be calculated.

The criticism of the assumptions underlying this approach can be illustrated in terms of Figure 1, which presents a simple structural model of the relationships among earnings, years of education, natural ability, and human capital. For these four variables, the figure depicts those which are determinants of human capital (the vertical arrows), of earnings (the horizontal arrows), and of both. Of the six links depicted, equation (1) includes only B, the relationship between years of education and earnings. The coefficient on S in (1) will be an unbiased estimate of the effect on earnings of skills acquired in school only under certain stringent conditions.

The first condition is that years of education must—through the relation F—provide an accurate measure of the human capital acquired in school. The market value of this human capital, determined by mar-

ginal product, must then determine earnings via A. However, years of education, being one input into the educational production function, may be a poor guide to the output from the function. Second, years of education must influence earnings only indirectly, through $F+A$. If there is a direct relationship, through B, which is positive, the coefficient on years of education overstates the human capital effect. The loose amalgam of hypotheses concerning the payment for educational qualifications irrespective of their economic value, generally known as "credentialism," stresses the direct effect of years of education on earnings. According to this view, schools provide students with a credential which is personally valuable but not productive. For instance, the government may determine wages and establish education-based hiring and payment criteria, or private employers may discriminate in favor of the educated with whom they share similar socioeconomic backgrounds.

The third condition is that, if ability is correlated with years of education (E), it must have no direct (C) or indirect (via $E+B$) effect on earnings: positive relationships imply that the coefficient on S in (1) overstates the effect on earnings of skills acquired in school, that is, the effect of ability is wrongly attributed to years of education. Employers may reward ability on an individual basis or, according to the theory of educational screening for ability, they may use years of education as a means of identifying workers who are potentially more productive, drawing on two stochastic relationships, E and C. Educational attainment "signals" workers with greater average ability, and it is this ability, rather than what is actually learned in school, that is rewarded. There is, however, a way in which ability can strengthen the human capital relationship between earnings and education. If educational selection criteria are meritocratic in the sense that they promote the more able (relation E), then years of education are more efficiently transformed into cognitive skills (on account of D): $E+F+A$ and $D+A$ influence but do not bias the estimate of the effect on earnings of skills acquired in school.

To capture the complex relationships depicted in Figure 1 we take cognitive skills to be a measure of human capital and reasoning ability to be a measure of predetermined natural ability, and posit a recursive model represented in the following three equations:

$$S = a_0 + a_1 R + a_2 E + a_{3i} F_i + v \tag{2}$$

$$H = b_0 + b_1 R + b_2 S + b_3 G + b_4 B + y \tag{3}$$

$$\ln W = c_0 + c_1 S + c_2 R + c_3 H + c_4 L + c_5 L^2 + z \tag{4}$$

where R = reasoning ability, E = an indicator of the aggregate probability of attending secondary school when the individual was aged 14, F_i = indicators of parental educational background, H = cognitive achievement, G = an indicator of attendance at a government (as opposed to private) school, B = an indicator of urban (as opposed to rural) birth, and v, y, z = disturbance terms.[5]

Equation (2) reflects the influence of natural ability on educational attainment within a subsidized and competitive educational system (relation E in Figure 1). Equation (3) is an educational production function, incorporating relations D and F; it is similar in form to those used in most such studies.[6] The earnings function specified in equation (4) includes relations A, B, and C. As opposed to the "conventional human capital model" in (1), we refer to (4) as the "expanded human capital model."

II. The Data

The data for this investigation come from rigorously comparable surveys administered by a team (including the authors) in the two countries within a few months of each other in 1980. The full samples were randomly

[5] Full definitions of variables are provided and the system is tested for recursiveness in the subsequent sections.

[6] See the reviews of Eric Hanushek (1979) and L. Lau (1979).

selected on an establishment basis, using a two-stage procedure, from among the wage-labor force of Nairobi and of Dar es Salaam. Establishments from all sectors of the urban economy—public and private, manufacturing and nonmanufacturing—are represented.[7] The main questionnaire was administered to the full sample, and provides the information on earnings,[8] education, employment experience, and other personal characteristics. Not all the respondents were given the three tests that yielded our measures of reasoning ability, literacy, and numeracy: testing was confined to a random subsample of primary- and secondary-completers.[9] Whereas the full samples each contained some 2,000 employees, the subsamples numbered 205 in Kenya and 179 in Tanzania. The analysis using the test results is necessarily confined to the subsamples. The survey design thus requires that schooling be entered as a dichotomous rather than a continuous variable.

Reasoning ability was tested by means of "Raven's Progressive Matrices" (see J. C. Raven, 1956). This test involves the matching of pictorial patterns, for which literacy and numeracy provide no advantage; it has been widely used in developing countries.[10] The tests of reading achievement and mathematics were designed by the Educational Testing Service of Princeton specifically for use in these surveys. The designs were based on questions in national primary-leaving and Form IV examinations and on other guides to the content of the academic curriculum, which is much the same in Kenya as in Tanzania. The major difference is that use of Swahili is stressed more in Tanzania; questions were therefore set in both English and Swahili for respondents to choose the language they preferred. The sum of the scores on the literacy and numeracy tests is used as the measure of cognitive skill.[11] The frequency distributions of test scores for each sample as a whole and for primary- and secondary-leavers reveal considerable variance on each test, a desirable characteristic for dependent and independent variable alike. All three tests appear to have been appropriate for the target groups: there are very few perfect scores and no zero scores, suggesting that the results do not suffer from the common problem of truncation of the ability or achievement distribution that arises when questions are too easy or too difficult.

Although these measures represent a distinct advance, their limitations should be recognized in the interpretation of the results to come. Because secondary schools select entrants partly on the basis of performance in primary-leaving examinations, the difference between primary- and secondary-leavers in mean achievement scores may exaggerate the value added by secondary education. However, our test of whether selection by cognitive achievement for secondary school qualifies our assumption of recursiveness proved negative. Noncognitive traits, such as attitudes and interests, may also be acquired in school and may be valued in the labor market. Natural ability may

[7]Neither very small establishments (< 5 employees) nor establishments in urban and rural areas outside the capital cities were included in our samples. Since selectivity by unmeasured personal characteristics in the surveyed capital city establishments is likely to be stronger at the primary than at the secondary level, however, any consequent sample selection bias is likely to understate the benefits of cognitive skill acquisition, so strengthening the argument of this paper.

[8]Sometimes referred to below as wages. As data are generally available only on earnings per month and not on hours worked, it is not possible to estimate separate wage and hours functions, nor to establish whether higher cognitive skill causes people to work more productively as opposed to longer hours. However, the influence of longer hours is unlikely to have been important: only 20 percent of the sample had worked overtime in Kenya, and 31 percent in Tanzania; the percentages for nonmanual employees being 10 and 23 percent, respectively. Whether overtime was worked appeared to depend mainly on the nature of the job and the characteristics of the employer.

[9]It took half an hour per respondent to complete the questionnaire and an hour per respondent to administer the tests. Given its small size, the subsample was stratified by education to ensure sufficient observations in strata of particular interest.

[10]See, for example, E. L. Klingelhofer (1967), M. Wober (1969), and U. Sinha (1968).

[11]The ability score is marked out of 36, the achievement score out of 63 (34 for numeracy and 29 for literacy).

involve not only reasoning power, but also such unmeasured but marketable qualities as drive, determination, and dynamism. If the test scores are too narrow as measures of natural ability and human capital formation, the R and H coefficients are likely to understate their importance. Insofar as the omitted variables are positively correlated with educational attainment, the coefficient on S is likely to overstate the importance of credentialism. Finally, the ability that we measure may not be due entirely to heredity and home environment: education may have enhanced reasoning power.[12] However, the fact that the weighted subsample mean values of R are not significantly different (27.8 in Kenya and 26.4 in Tanzania), whereas those of H are significantly different (40.0 and 30.3, respectively), on account of the greater quantity and quality of secondary education in Kenya,[13] suggests that R is not acquired in school.

III. The Expanded Human Capital Earnings Function

Estimates of the conventional human capital earnings function (equation (1)) for Kenya and Tanzania are shown in column 1 of Table 1.[14] In Kenya workers are paid a premium of 4.2 percent per year of employment experience, and secondary-leavers are paid 61 percent more than primary-leavers. In Tanzania the returns to experience are higher (5.5 percent), but because of Tanzania's vigorously imposed pay policy the gain from secondary education, though substantial, is lower (32 percent).[15]

A. *Do Cognitive Skills and Ability Matter?*

Column 2, Table 1, permits a comparison of estimates of the conventional and expanded human capital earnings functions in Kenya and Tanzania. In neither country are the estimated returns to experience affected by the introduction of variables measuring (cognitive) achievement and (reasoning) ability. By contrast, the premium on secondary education declines by nearly two-thirds in both countries, and in Tanzania it is no longer significantly different from zero. In neither country is the independent influence of ability on earnings either large or significant. By contrast, in both countries the coefficient on the achievement score is positive, significant, and large relative to the coefficient on the ability score.[16]

B. *Do Cognitive Skills Matter for Manual as Well as for Nonmanual Workers?*

The results of the stratified regressions (cols. 5 and 6) show that in both countries the payment for cognitive skills is not confined to white-collar workers: manual workers are also rewarded for literacy and numeracy. Although the coefficient on H is higher for nonmanual (0.017) than for manual workers (0.013) in Kenya and also in Tanzania (0.012 and 0.008, respectively), F-tests indicate that in neither country is the difference in the coefficient on H significant as between occupations. It seems that accomplishment in the basic skills of reading and numbering enables mechanics, machinists, and fork-lift drivers as well as accountants, clerks, and secretaries to do a better job.[17] By contrast, in no case is the coefficient on R significant.[18]

[12] In that case, the coefficient on ability is liable to be upward biased and that on achievement downward biased.

[13] See Knight and Sabot (1984).

[14] The equations were also estimated with a squared experience term, but whereas the coefficient was negative as expected, it was not significantly different from zero.

[15] The Tanzania government has compressed the structure of wages in the dominant, public sector. In the relatively unfettered private sector, the premium on secondary education is higher than in the public sector and, indeed, higher than in Kenya. We recognize that the competitive market value of secondary education in Tanzania is greater than our estimates suggest. See Knight and Sabot (1983).

[16] This result holds when either the literacy or numeracy score replaces the combined score.

[17] In only one of the four cases (manual workers in Tanzania) is the coefficient on achievement not significant at the 5 percent level.

[18] When the samples are stratified instead by educational levels, F-tests indicate precisely equivalent results for the achievement variable. The effect of ability on earnings remains small by comparison with the effect on achievement, and not significantly different from zero in three of the four cases.

TABLE 1—HUMAN CAPITAL EARNINGS FUNCTIONS WITH AND WITHOUT MEASURES OF ABILITY AND COGNITIVE ACHIEVEMENT[a]

	Whole Subsample (1)	Whole Subsample (2)	Primary Leavers (3)	Secondary Leavers (4)	Manual Workers (5)	White-Collar Workers (6)
Kenya						
L	.042	.045	.031	.062	.036	.049
	(8.40)	(9.84)	(4.49)	(10.20)	(6.02)	(8.64)
S	.476	.132	–	–	.065	.030
	(6.70)	(2.[illegible])			(0.650)	(0.23)
H	–	.020	.019	.023	.013	.017
		(6.[illegible])	(3.98)	(5.40)	(3.21)	(3.55)
R	–	.006	−.000	.014	.003	.011
		(1.[illegible])	(0.02)	(2.17)	(0.50)	(1.46)
Constant	6.297	5.[illegible]59	5.811	5.171	5.866	5.705
R^2	.29	.[illegible]	.39	.50	.32	.49
N	205	205	71	134	116	88
Tanzania						
L	.054	.055	.049	.066	.044	.061
	(9.70)	(10.[illegible]0)	(7.13)	(7.06)	(4.88)	(7.82)
S	.280	.012	–	–	.141	.068
	(4.30)	(1.[illegible]2)			(0.85)	(0.58)
H	–	.013	.009	.013	.008	.012
		(3.[illegible]2)	(1.66)	(2.29)	(1.16)	(2.25)
R	–	.001	−.001	.010	−.004	.013
		(0.[illegible]5)	(0.21)	(1.01)	(0.64)	(1.51)
Constant		5.[illegible]52	5.908	5.476	5.027	5.423
R^2	.38	.[illegible]3	.34	.47	.24	.46
N	179	179	107	72	87	88

[a]The dependent variable is $\ln W$. The t-statistics are shown in parentheses.

C. *Could Cognitive Skills Represent Anything But Human Capital?*

Administered wage scales might explain why employers would pay a premium to workers with more years of education even if they were not more productive. Screening for ability might similarly explain such a premium even if the cognitive skills acquired in school had no economic value. Neither of these accounts, however, could also explain why cognitive skills are rewarded within an educational stratum.

Whereas employers could ascertain the length of education of job applicants, they did not have our test scores to provide them with independent measures of numeracy and literacy. Grades of pass in the national secondary-leaving examination do provide employers with a ready indication of cognitive achievement and ability. There is evidence for Kenya, where competition for jobs among secondary-leavers is intense, that examination scores are used as a selection criterion. We therefore expect, and find, a significantly positive relation between grade of pass and starting wage.[19] Similarly, our achievement test score bears a positive and significant relationship to the starting wage in Kenya. If, however, these results reflected the favoring of good examinees for reasons of "fairness" or for screening purposes, rather than simply for their cognitive skills, we would expect the relation to decline as employment experience lengthens. On the contrary, in Kenya, achievement as measured by scores on our test is a markedly better predictor of current than of starting wages.[20]

[19]In a Kenya earnings function for secondary-leavers in which the worker's (constant price) starting wage is the dependent variable.

[20]The substitution of ln starting for ln current wage as the dependent variable in column 2 of Table 1 results in a reduction in the coefficient on H from .020 to .011,

In Tanzania, where secondary-leavers are in scarcer supply, there is no significant relation between starting wage and grade of secondary school pass or achievement score.[21] Yet the current returns to cognitive achievement for secondary-levers are positive and significant in both countries. Whereas in neither country do employers have ready equivalent measures of the cognitive skills of primary-leavers, in both countries the returns to cognitive achievement are of the same order of magnitude to primary- and to secondary-leavers. It would seem that employers discover the cognitive skills of their workers on the job, and that they are willing to pay for these skills.

D. *Why Do Secondary-Leavers Earn More than Primary-Leavers?*

The coefficients on the independent variables can only be suggestive of their relative importance. This, and subsequent exercises, provide measures of the relative effects of the independent variables in the earnings function on the structure and dispersion of earnings. The gross difference in (geometric) mean wages (G) between primary- and secondary-leavers (24 percent in Kenya and 30 percent in Tanzania) is decomposed.[22] The earnings of primary-leavers (denoted by the subscript p) are determined by the earnings function for primary-leavers, and by their characteristics, represented by the vector $Z_p: W_p = F_p(Z)_p$; similarly, $W_s = F_s(Z_s)$ where s denotes secondary-leavers. A bar indicates the mean value of a variable:

$$(5)\quad G = \overline{W}_s - \overline{W}_p = F_s(\overline{Z}_s) - F_p(\overline{Z}_p)$$
$$= F_s(\overline{Z}_s - \overline{Z}_p) + \left(F_s(\overline{Z}_p) - F_p(\overline{Z}_p)\right).$$

The former is the component "explained" by the differences in the mean characteristics of the two groups, and the latter is the "unexplained" component which results from differences in the constant term and coefficients of the earnings functions.[23]

We simulate the effect on the predicted wage of a representative primary-leaver (with the mean characteristics of his group) of imposing, each in turn, the characteristics of a representative secondary-leaver. In the case of achievement, for instance, the effect is to raise $\overline{W}_p$ in the proportion $c_{3p}(\overline{H}_s - \overline{H}_p)$. The effect of the difference in length of education is obtained from the unexplained residual in (5), which reflects group differences in earnings functions. The relative contributions to this premium that are made by group differences in cognitive skills, ability, years of education, and employment experience are shown in Table 2.

Secondary-leavers do not earn more because of differential experience on the job: they have *less* experience than primary-leavers, markedly so in Kenya and marginally so in Tanzania.[24] Nor does the small difference in ability as between the two educational groups explain why secondary-leavers earn more. The direct returns to ability are so low that giving primary-leavers the ability levels of secondary-leavers would increase their earnings by some 0–7 percent in Kenya and by 0–4 percent in Tanzania.[25] Giving primary-leavers four more years of education would, *ceteris paribus*, substantially increase their earnings, by 15–24 percent in Kenya and 8–18 percent in Tanzania. This could be a reflection of credentialism or of screening, but it could alternatively be the result of unmeasured noncognitive skills acquired in secondary education. The largest

i.e., the percentage response of current wage to a unit increase in H is nearly twice as great as that of starting wage.

[21] Using the same specification for Tanzania as for Kenya.

[22] By means of a technique taken from the literature on labor market discrimination; Alan Blinder (1973) and Ronald Oaxaca (1973) are pioneering examples. We decompose the differences in geometric mean wages, i.e., in antilog mean $\ln W$, because the earnings function has $\ln W$ as dependent variable.

[23] Alternatively, the decomposition can be based on $F_p(Z_s)$ instead of $F_s(Z_p)$.

[24] Reflecting not only their later entry to the labor force but also the expansion of secondary education, and the more rapid expansion in Kenya than Tanzania.

[25] In each case, the lower end of the range is the estimate yielded by the earnings function for primary-leavers, whose returns are generally lower, and the upper end is that yielded by the earnings function for secondary-leavers.

TABLE 2—THE EFFECT OF INTRODUCING THE CHARACTERISTICS OF A REPRESENTATIVE SECONDARY-LEAVER ON THE PREDICTED WAGE OF A REPRESENTATIVE PRIMARY-LEAVER

	Mean Value		Change in Predicted Wage Using:					
			Primary-Leaver Coefficients			Secondary-Leaver Coefficients		
	Secondary-Leavers	Primary-Leavers		$\Delta \overline{W}_s$			$\Delta \overline{W}_p$	
	$\overline{Z}_s$	$\overline{Z}_p$	$\Delta \ln \overline{W}_s$	Shillings	Percent	$\Delta \ln \overline{W}_p$	Shillings	Percent
Kenya								
W	1141.0	918.0						
$\ln W$	7.040	6.822						
H	46.3	32.3	.266	280	30.5	.322	349	38.0
R	30.3	25.7	.000	0	0.0	.064	61	6.6
L	6.4	12.6	−.192	−194.6	−21.2	−.384	−429	−46.8
S	–	–	.143	141.3	15.4	.215	220	24.0
Tanzania								
W	843.0	649.0						
$\ln W$	6.737	6.475						
H	37.5	24.7	.115	79	12.2	.166	117	18.1
R	29.0	25.0	−.004	−3	−0.4	.040	27	4.1
L	7.2	7.5	−.015	−10	−1.5	−.020	−13	−2.0
S	–	–	.165	116	17.9	.075	51	7.8

Notes: The change in the predicted geometric mean wage of primary- or secondary-leavers as the result of the addition or subtraction of four years of secondary education is derived as a residual (the remaining difference in geometric mean wages of the two groups) after eliminating the differences due to differences in the mean characteristics.

The percentage change in the geometric mean wage is calculated from the change in $\ln \overline{W}$ in a way analogous to the dummy variable in semilogarithmic earnings functions explained by Halvorsen and Palmquist.

The differences between primary- and secondary-leavers in the mean values of H and R are significant at the 1 percent level in both countries.

$\Delta \ln \overline{W}_s = F_p(\overline{Z}_s - \overline{Z}_p)$ and $\Delta \ln \overline{W}_p = F_s(\overline{Z}_s - \overline{Z}_p)$.

increase in wages would result from giving primary-leavers the higher achievement levels of secondary-leavers: 31–38 percent in Kenya and 12–18 percent in Tanzania.

E. *Do High-Achieving Primary-Leavers Earn More than Low-Achieving Secondary-Leavers?*

Columns 1 (Kenya) and 5 (Tanzania) of Table 3 show substantial variation in cognitive development within educational strata. The average achievement test score of the top third of primary-leavers is double that of the bottom third in both countries. Among secondary-leavers, the average score of the top third is half as much again as the bottom third in Kenya, and double that of the bottom third in Tanzania. In both countries, the literacy and numeracy of the top third of primary-leavers is roughly equal to that of the middle third of secondary-leavers.[26]

To estimate the impact on earnings within each educational group of the within-group variance of cognitive achievement, the estimated stratified earnings functions are used to predict earnings for different levels of cognitive achievement. In the case of primary-leavers:

$$(6)\quad \left(\widehat{\ln W_{PH}}\right)_i = c_{0p} + c_{2p}\overline{R}_p + c_{4p}\overline{L} + c_{3p}H_{pi}$$

$$(7)\quad \left(\widehat{\ln W_{PR}}\right)_i = c_{0p} + c_{3p}\overline{H}_p + c_{4p}\overline{L} + c_{2p}R_{pi}$$

where H_{pi} and R_{pi} represent the achieve-

[26] It seems that cognitive skills are not the only basis for access to secondary education.

TABLE 3—PREDICTED WAGES OF PRIMARY- AND SECONDARY-LEAVERS WITH VARYING LEVELS OF COGNITIVE ACHIEVEMENT AND REASONING ABILITY

		Kenya				Tanzania			
		By Achievement		By Ability		By Achievement		By Ability	
		$\bar{H}_i$ (1)	$\bar{Y}_i$ (2)	$\bar{R}_i$ (3)	$\bar{Y}_i$ (4)	$\bar{H}_i$ (5)	$\bar{Y}_i$ (6)	$\bar{R}_i$ (7)	$\bar{Y}_i$ (8)
Primary-Leavers									
Bottom	10%	13.1	532	10.9	807	11.9	571	9.1	657
Bottom	1/3	21.4	623	16.0	806	16.8	598	16.7	651
Middle	1/3	31.2	751	26.0	804	24.6	643	26.1	643
Top	1/3	45.0	978	32.2	803	32.0	689	31.1	639
Top	10%	51.6	1,109	34.0	803	40.7	747	33.4	637
Secondary-Leavers									
Bottom	10%	28.1	864	17.4	1,083	20.1	681	13.4	732
Bottom	1/3	36.1	1,036	24.2	1,196	25.6	725	21.4	792
Middle	1/3	47.2	1,333	31.5	1,323	37.3	847	29.9	862
Top	1/3	54.0	1,556	34.9	1,387	48.5	983	33.8	896
Top	10%	55.9	1,624	35.3	1,395	52.6	1,039	35.5	911

ment and ability scores of each primary-leaver i and a circumflex indicates a predicted value.

Columns 2 and 6 show the predicted mean wages of primary- and secondary-leavers of varying levels of achievement but of the same levels of ability and experience. Secondary-leavers who scored in the top third on the achievement test are predicted to earn some 50 percent more than those in the bottom third in Kenya, and some 35 percent more in Tanzania; roughly the same percentages apply to primary-leavers. In both countries, it would seem, how much you learn in primary or in secondary school has a substantial influence on your performance at work. Moreover, the predicted wage of primary-leavers who scored in the top third is nearly as high as that of secondary-leavers who scored in the bottom third. In East Africa, mere attendance at secondary school is no guarantee of success in the labor market: it is necessary to learn one's school lessons.

F. *Do More Able Primary-Leavers Earn More than Less Able Secondary Leavers?*

There is substantial variation in reasoning ability within the two educational strata (cols. 3 and 7, Table 3). As with achievement, the ability of the top third of primary-leavers is roughly equal to that of the middle third of secondary-leavers. In contrast to variation in achievement, however, variation in ability has no effect on the predicted earnings of primary-leavers and little on those of secondary-leavers. Moreover, whereas the ability scores of the ablest 10 percent of primary-leavers are roughly double those of the least able 10 percent of secondary-leavers, their predicted wages are lower (cols. 4 and 8). In neither country is being among the most able of your peers a sufficient condition for successful performance in the labor market.

G. *How Much Inequality is Due to Cognitive Skills?*

The effects of ability, cognitive development, or years of education on the dispersion of earnings may differ in relative importance from their effects on the structure of earnings. The latter depends only on the size of the coefficients in the earnings function. The former depends also on the proportion of employees with a particular characteristic (in the case of the dummy variable), or the extent to which employees differ in possession of that characteristic (in the case of

TABLE 4—THE RELATIVE CONTRIBUTIONS OF WORKER CHARACTERISTICS TO THE DISPERSION OF EARNINGS; THE MEAN CHARACTERISTICS OF WORKERS BY EARNINGS QUINTILE

	Kenya			Tanzania		
Contribution to Variance:	Absolute	Percentage of Total	Percentage of Restricted Total	Absolute	Percentage of Total	Percentage of Restricted Total
L	.031	32.0	–	.095	72.0	–
S	.011	11.3	16.7	.011	8.3	29.8
H	.049	50.5	74.2	.025	18.9	67.6
R	.006	6.2	9.1	.001	.8	2.7
Total	.097	100.0	100.0	.132	100.0	100.0

Earnings Quintile:	Kenya Lowest	Second	Third	Fourth	Highest	Tanzania Lowest	Second	Third	Fourth	Highest
$\bar{L}$	6.45	8.64	6.62	7.73	13.28	3.16	5.73	7.68	8.69	12.07
$\bar{S}$	.43	.45	.78	.70	.74	.23	.38	.38	.32	.61
$\bar{H}$	31.69	38.00	44.60	43.82	46.61	26.00	28.98	29.43	27.56	36.16
$\bar{R}$	25.32	27.57	29.71	29.80	29.21	25.41	25.14	27.60	26.10	27.54

continuous variables); and where in the distribution of pay those who possess the characteristic or possess it in varying degrees are found.

To measure relative contributions to dispersion, we adopt the following procedure: using equation (4), written here as $\ln W_i = a + \Sigma_j b_{ij} Z_{ij}$, where Z_{ij} is the set of independent variables $(j = 1 \dots n)$, we predict the earnings of each employee $(\hat{W}_i)$. Each independent variable (j) is in turn set equal to its mean value, and predicted earnings $(\hat{W}_{ij})$ are estimated using the set of other characteristics possessed by each employee. Here $\hat{W}_{ij}$ represents the predicted value of W for each individual (i) when his endowment of j equals that of all other individuals. The variances of $\hat{W}_i$ and $\hat{W}_{ij}$ are calculated, and the contribution of Z_j to the explained variance of earnings is estimated as $\mathrm{var}(\hat{W}_i) - \mathrm{var}(\hat{W}_{ij})$. The relative contribution of each individual variable is calculated by expressing its contribution as a percentage of $\Sigma_j(\mathrm{var}(\hat{W}_i) - \mathrm{var}(\hat{W}_{ij}))$.[27] In effect we are attempting to answer the following counterfactual question: what would be the effect on the inequality of pay if, while mean earnings were held constant, the dispersion due to a particular characteristic such as cognitive achievement was eliminated?

The relative contribution to inequality of each independent variable in the expanded human capital earnings function for the unstratified sample is shown in Table 4. The contribution of employment experience to the variance of earnings is markedly greater in Tanzania than in Kenya.[28] The contribution of the ability variable to the variance of earnings is small in both countries, partly because of its negligible coefficient and partly because high and low earners have similar ability scores. The contribution of years of education is larger, reflecting the size of its coefficient and the tendency for the proportion with secondary education to rise with earnings quintile. In Kenya achievement accounts for three-quarters of the variance in earnings explained jointly by ability, education, and achievement; in Tanzania the share is two-thirds. Not only are cognitive skills highly rewarded, but there are few highly literate and numerate workers, be they

[27]For further explication of this method of decomposing inequality and a comparison with other methods, see Behrman, Knight, and Sabot (1983).

[28]In Tanzania mean experience rises monotonically, from 3.1 years in the lowest to 12.1 years in the highest earnings quintile. This is not the case in Kenya; i.e., high levels of experience are associated with low as well as with high incomes, possibly because of the inverse correlation between education and experience: the more educated, who are more plentiful in Kenya, have received preference over the more experienced but less educated in access to jobs.

primary- or secondary-leavers, in the low-earnings quintiles.

IV. The Educational Production and Attainment Functions and Indirect Effects on Earnings

Having shown that length of education has a relatively small and ability a negligible direct influence on earnings, we now examine a possible indirect influence through their effects on cognitive achievement. The simple correlations between S and H and between R and H are strong and positive. The mean achievement scores are significantly higher for secondary- than for primary-leavers (43 percent higher in Kenya and 52 percent in Tanzania), and there is a monotonic relationship between ability groups and their mean levels of achievement.

An educational production function, based on equation (3), is presented in Table 5 using a linear specification.[29] In both countries, cognitive achievement bears a highly significant positive relationship to educational level and to ability. In Kenya, secondary education raises H by 11.75 points, or by 35 percent at the mean; very similar results are obtained for Tanzania. The elasticity of response of cognitive skill to reasoning ability at the mean is roughly 0.4 in both countries. In Kenya the coefficient on G (a dummy variable taking a value of 1 if the secondary school attended by a secondary-leaver, and the primary school attended by a primary-leaver, was a government school, and 0 otherwise) is significantly positive. In both countries the coefficient on B (a dummy variable indicating birth in an urban area, birth in a rural area being the omitted category) is almost significantly negative.[30]

TABLE 5—EDUCATIONAL PRODUCTION FUNCTIONS

Variable	Kenya	Tanzania
S	11.754	10.939
	(8.50)	(8.84)
G	3.366	0.995
	(2.49)	(0.76)
B	−3.567	−2.651
	(1.78)	(1.82)
R	0.560	0.487
	(5.55)	(5.58)
Constant	15.49	12.34
$\overline{H}$	39.98	30.33
R^2	0.42	0.44
Standard Error	8.77	7.76
Percentage Standard Error	21.1	26.2

Notes: The dependent variable is H; t-statistics are shown below the coefficients in parentheses; the mean values of variables, here and elsewhere, are derived from the subsamples weighted according to the proportions in which primary- and secondary-completers are found in the full samples.

An educational attainment function, based on equation (2), was estimated by means of probit analysis. The results are very similar in the two countries, being

$$\hat{p}=\phi(\underset{(4.051)}{-1.816}+\underset{(3.075)}{.049R}+\underset{(3.918)}{.070E}+\underset{(.752)}{.184F_1}+\underset{(1.975)}{.530F_2})\qquad \chi^2=46.54$$

$$\hat{p}=\phi(\underset{(3.357)}{-1.760}+\underset{(3.889)}{.067R}-\underset{(2.484)}{.248E}+\underset{(.515)}{.133F_1}+\underset{(3.426)}{.929F_2})\qquad \chi^2=30.92$$

in Kenya and Tanzania, respectively, where $\hat{p}$ is the probability of going on to secondary school, E is the number of secondary school places as a proportion of the number of 14-year olds when the respondent was aged 14, F_1 indicates that one parent and F_2 that both had received education, $\phi(\cdot)$ denotes the cumulative unit normal distribution, and

[29] A log-linear specification (with the continuous variables H and R in natural logarithms) was also estimated but was inferior in terms of the percentage standard error of H (29 percent in Kenya and 31 percent in Tanzania) and the significance of some coefficients. The ensuing simulation analysis is based on the linear specification but the results are not sensitive to the choice of specification.

[30] This counterintuitive result may reflect greater selectivity in access to schooling and to the urban labor market among the rural born.

the figures in parentheses are t-statistics. The probability of going on to secondary school was positively and significantly related to the ability score, it was raised significantly if both parents had been educated, and it was significantly affected by the secondary enrollment ratio—positively, as expected, in Kenya but negatively in Tanzania. The reason for this negative sign is that although E rose over time, the proportion of primary school completers continuing to secondary school actually fell. With all independent variables at their mean values, an increase in the ability score from the mean of the bottom to that of the top-ability tercile would raise the probability of secondary school attendance by .25 in Kenya and by .35 in Tanzania. Ability therefore has two indirect effects on earnings: not only via relation D but also via relation E in Figure 1.

Before combining the three functions for simulation analysis, we test whether the estimated model is recursive; that is, whether the estimates are consistent and not subject to simultaneous equation bias. If some unmeasured characteristics, such as drive and determination, contributed to educational attainment, to cognitive achievement, and to earnings, the error terms (v, y, and z, respectively) in equations (2), (3), and (4) would be correlated, as would educational attainment and y, educational attainment and z, and cognitive skill and z. Applying a specification test developed by Jerry Hausman (1978), we added the predicted value of educational attainment ($\hat{S}$) for each individual as an independent variable in (3) and in (4), and the predicted value of cognitive skill ($\hat{H}$) as an independent variable in (4).[31] Our findings that the coefficients are not significantly different from zero in five of the six cases and just significantly so in the sixth makes it difficult to reject the null hypothesis that the equation system is recursive.[32]

A further test of recursiveness between equations (2) and (3) was conducted. Equation (3) was estimated using instrumental variables, and the estimated coefficients were used to generate $\hat{H}$ for each individual at the end of primary school (i.e., with $S=0$). Equation (2) was then estimated with $\hat{H}$ as an additional independent variable. The coefficient on $\hat{H}$ is not significant in either country suggesting that simultaneity on account of selection for secondary school by cognitive achievement is unlikely.[33]

The indirect effects of ability are measured and compared with the direct effect in Table 6. Two ability levels are considered in each sample, corresponding to the mean values of R for the top- and bottom-ability terciles; all other characteristics of the sample are kept at their mean values. Within the three-equation system, we then trace the difference in predicted wages between the two ability levels which is due to relations C, D, and E in Figure 1. The full consequence of the assumed ability difference—incorporating all three effects—is to create a difference in predicted wages equal to 32 percent of the sample mean in Kenya and 16 percent in Tanzania (the final row of the table). The direct effect of ability differences on earnings, working through the earnings function alone, accounts for only one-fifth of the predicted full wage difference in Kenya and for much less in Tanzania (relation C). The indirect effect of ability on cognitive skill acquisition and hence on earnings represents

[31] $\hat{S}$ and $\hat{H}$ are generated using (2) and (3), respectively, plus the other exogenous variables in the three-equation system.

[32] The coefficients are -1.518 (t-value $=.373$), $-.058$ (.210), and $-.011$ (.748) in Kenya, and $-.243$ (1.430), .433 (2.036), and $-.011$ (.647) in Tanzania The possibility of simultaneity between equations (2) and (4) in Tanzania makes the Tanzanian results less reliable. However, the fact that the coefficient on $\hat{S}$ is significantly positive implying that the coefficient on S is biased downwards, suggests that the bias is not due to simultaneity. The extreme rationing of secondary enrollment in Tanzania ensures that the private demand remains strong—as revealed by the private rate of return and subjective responses to survey questions—despite government compression of the earnings structure (fn. 17). The suggestion that the significant positive coefficient is due to the less perceptive acquiring unprofitable education and receiving lower income, is therefore implausible.

[33] The coefficient was actually negative, being -0.090 (standard error $=0.053$) in Kenya, and -0.063 (0.079) in Tanzania.

TABLE 6—THE DIRECT AND INDIRECT EFFECTS OF ABILITY ON EARNINGS

Mean Values for the Top and Bottom Terciles Classified by Reasoning Ability	Kenya				Tanzania			
			Difference				Difference	
	Top Tercile	Bottom Tercile	Absolute	As Percentage of Total[a]	Top Tercile	Bottom Tercile	Absolute	As Percentage of Total[a]
R	33.8	19.8	14.0		33.3	18.5	14.8	
p	0.77	0.52	0.25		0.54	0.19	0.35	
H (all effects):	45.8	34.9	10.9		34.7	23.7	11.0	
Predicted Wages Showing the Effect of:[b]								
Relation C	1,064	979	85	24	749	739	10	7
Relation D	1,089	929	160	44	772	704	68	48
Relation E	1,127	1,012	115	32	765	700	65	45
Relation E' (human capital only)	1,076	1,014	62	17	754	718	36	25
Relations C, D, E (all effects)	1,250	890	360	100	804	661	143	100

[a]As the sum of the separate effects is not exactly equal to their combined effect, each is expressed as a percentage of the sum.

[b]The measure of each effect is derived from equations (2), (3), and (4). It shows the effect on the wage of replacing the mean value of ability for the subsample ($\overline{R}$) by the mean value for the upper or lower tercile ($\overline{R}_t$). The multiplicands are as follows: Relation C: c_2; Relation D: $c_3 \cdot b_1$; Relation E: $c_1 \cdot p(a_1) + c_3 \cdot p(a_1) \cdot b_2$; Relation E': $c_3 \cdot p(a_1) \cdot b_2$; Relations C, D, E: $c_2 + c_3(b_1 + p(a_1) \cdot b_2) + c_1 \cdot p(a_1)$ (for instance, in the case of relation C: $(\overline{\ln W})_t - (\overline{\ln W}) = c_2(\overline{R}_t - \overline{R})$).

38 percent in both countries (relation D), and the indirect effect of ability via educational attainment about a third (relation E). At least half of this effect works through human capital acquisition (relation E') rather than credentialism.

It is also possible to distinguish the different effects of secondary school attendance on earnings. The directly observed effect (relation B in Figure 1) is derived from the coefficient c_1 in equation (4). The value is 0.19 in Kenya and 0.11 in Tanzania, implying that the wage is raised by 21 and 12 percent, respectively, by what we termed credentialism. The other effect (relations F and A in Figure 1) is derived from a combination of equations (3) and (4). The coefficient b_2 in the former shows the effect of secondary schooling on cognitive skills, and c_3 in the latter the effect of cognitive skills on earnings. Their product $b_2 \cdot c_3$ (0.22 in Kenya and 0.14 in Tanzania) indicates that human capital acquisition in secondary school raises earnings by 25 and 15 percent, respectively. The human capital effect of a secondary education thus exceeds the credentialist effect. In summary, use of the three-equation system has shown that the indirectly measured effects of differences in reasoning ability and in educational attainment both exceed the direct effects.

V. Conclusions

Our survey data from East Africa have permitted a sharper test than hitherto of the competing explanations—credentialism, ability, screening, or human capital—of why workers with secondary education earn more. The direct returns to reasoning ability in the labor market are small, those to years of education are moderate, and those to literacy and numeracy—dimensions of human capital—are large. The returns to cognitive achievement are not significantly lower for manual than for nonmanual workers.

The returns to cognitive skills cannot but be a payment for human capital. The direct

returns to years of education, on the other hand, could reflect credentialism or screening or human capital acquired at school or at home; that is, their interpretation is inconclusive. It appears that literate and numerate workers are more productive, and that education is valuable to workers because it can give them skills that increase their productivity. These conclusions have generally satisfied the usual statistical tests. Their robustness derives no less from the fact that they all apply to both Kenya and Tanzania.

The main effects of length of education and reasoning ability on earnings are indirect, operating through the development of cognitive skills. More educated or brighter workers tend to be more literate and numerate. The main reason why secondary-leavers earn more on average than primary-leavers is their higher average level of cognitive achievement. However, there is substantial variation in cognitive achievement, and also in reasoning ability, within the two educational groups. Whereas primary-leavers of high ability earn less than less able secondary-leavers, this is generally not the case for cognitive skills. Within each educational group, high achievers earn a great deal more than low achievers. Just as cognitive achievement is the main determinant of the structure of earnings, so also—far more than reasoning ability and school attendance—does it account for much of the inequality of earnings among workers. Because inequality is primarily due to differences in productivity based on cognitive skills, the efficiency cost of reducing inequality may be high.

Our analysis provides strong support for the human capital interpretation of the educational structure of wages. Whether these conclusions should be generalized beyond East Africa to the many other countries in which rates of returns have been estimated is, however, open to question. Kenya and Tanzania have much lower incomes, and cognitive skills are in shorter supply, than in most developing counties, particularly those of Asia and Latin America. As economic development proceeds, the growth of educated labor may outstrip the growth of the economy. In that case, the returns to cognitive achievement may decline, while for political and institutional reasons the returns to years of education may remain high.

REFERENCES

Arrow, Kenneth J., "Higher Education as a Filter," *Journal of Public Economics*, July 1973, *2*, 193–216.

Behrman, Jere R. et al., *Socioeconomic Success: A Study of the Effects of Genetic Endowments, Family Environment and Schooling*, Amsterdam: North-Holland, 1980.

Behrman, Jere, Knight, J. B. and Sabot, R. H., "A Simulation Alternative to the Comparative R^2 Approach to Decomposing Inequality," *Oxford Bulletin of Economics and Statistics*, August 1983, *45*, 307–12.

Blaug, Mark, "Human Capital Theory: A Slightly Jaundiced View," *Journal of Economic Literature*, September 1976, *14*, 827–55.

Blinder, Alan S., "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, Fall 1973, *8*, 436–55.

Bowles, Samuel and Gintis, Herbert, *Schooling in Capitalist America*, London: Routledge and Kegan Paul, 1976.

Chamberlain, Gary and Griliches, Zvi, "More on Brothers," in Paul Taubman, ed., *Kinometrics: Determinants of Socioeconomic Success within and between Families*, Amsterdam: North-Holland, 1977.

Griliches, Zvi, "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, January 1977, *45*, 1–22.

______, "Sibling Models and Data in Economics: Beginnings of a Survey," *Journal of Political Economy*, October 1979, Suppl., *87*, S37–64.

______ **and Mason, William M.**, "Education, Income, and Ability," *Journal of Political Economy*, May/June 1972, Suppl., *80*, S74–103.

Halvorsen, Robert and Palmquist, Raymond, "The Interpretation of Dummy Variables in Semilogarithmic Equations," *American Economic Review*, June 1980, *70*, 474–75.

Hanushek, Eric A., "Conceptual and Empirical

Issues in the Estimation of Educational Production Functions," *Journal of Human Resources*, Summer 1979, *14*, 351–88.

Hausman, Jerry, "Specification Tests in Econometrics," *Econometrica*, November 1978, *46*, 1251–71.

Klingelhofer, E. L., "Performance of Tanzanian School Pupils on the Raven's Matrices Test," *Journal of Social Psychology*, August 1967, *72*, 205–16.

Knight, J. B., and Sabot, R. H., "Educational Expansion, Government Policy and Wage Compression," The World Bank, 1983.

______ **and** ______, "Educational Policy and Labor Productivity: An Output Accounting Exercise," The World Bank, 1984.

Lau, L., "Educational Production Functions," in Report of a Committee of the National Academy of Education, *Economic Dimensions of Education*, Washington, 1979.

Oaxaca, Ronald, "Male-Female Differentials in Urban Labor Markets," *International Economic Review*, October 1973, *3*, 673–709.

Olneck, Michael R., "On the Use of Sibling Data to Estimate the Effects of Family Background, Cognitive Skills and Schooling: Results from the Kalamazoo Brothers Study," in Paul Taubman, ed., *Kinometrics: Determinants of Socioeconomic Success within and between Families*, Amsterdam: North-Holland, 1977.

Psacharopoulos, George (assisted by Keith Hinchcliffe), *Returns to Education. An International Comparison*, Amsterdam: Elsevier, 1973.

______, "Returns to Education: An Updated International Comparison," *Comparative Education*, March 1981, *17*, 321–41.

Raven, J. C., *Guide to the Coloured Progressive Matrices (Sets A, Ab, B)*, London: H. K. Lewis and Company, 1956.

Riley, John G., "Testing the Educational Screening Hypotheses," *Journal of Political Economy*, October 1979, Suppl., *87*, S227–52.

Sinha, U., "The Use of Raven's Progressive Matrices Test in India," *Indian Educational Review*, January 1968, *3*, 118–32.

Spence, Michael, "Competition in Salaries, Credentials, and Signalling Prerequisites for Jobs," *Quarterly Journal of Economics*, February 1976, *90*, 51–74.

Taubman, Paul, *Sources of Inequality of Earnings*, Amsterdam: North-Holland, 1975.

______ **and Wales, Terence,** *Higher Education and Earnings: College as an Investment and as a Screening Device*, New York: McGraw-Hill, 1974.

Thurow, Lester C., *Generating Inequality*, New York: Basic Books, 1975.

Wise, David A., "Academic Achievement and Job Performance," *American Economic Review*, September 1975, *65*, 350–66.

Wober, M. "The Meaning of Stability of the RPM Among Africans," *International Journal of Psychology*, 1969, *4*, 310–24.

The Labor Contract and True Economic Pension Liabilities

By RICHARD A. IPPOLITO*

Most pension-covered workers in the United States are covered solely or primarily by defined benefit plans.[1] In these plans, workers are typically promised a pension that is proportional to their years of service in the firm and their final wage. Depending upon how workers pay for it, the existence of a pension in the labor contract may play either an important role or no role at all in the worker-firm relationship. This paper addresses the issue, how workers pay for the cumulating pension benefit over their tenure cycle in the firm, and derives several economic implications for the worker-firm contract.

While the theory itself has ambiguous implications, the data confirm the longstanding presumption that pensions offer strong incentives for the worker to complete the normal tenure cycle in the firm, not to quit early. But if this intuitive notion is accepted, it is shown that a perhaps less intuitive notion is also true: firms make workers long-term bondholders in the firm. The same flow of worker pension savings that tie workers to the firm are not accounted for in the pension trust fund: pension plans are significantly underfunded in a true economic sense. This is a peculiar characteristic of the labor contract which raises several unanswered questions. The paper also shows that the nature of the pension contract can play an important role in determining the wage-service profile and the capitalized value of the firm.

Recent contributions to the literature pose an interesting and unconventional view of pensions, one which leads to the conclusion that pensions play a relatively minor role in the worker-firm contract. The argument has been implicitly made by Burt Barnow and Ronald Ehrenberg (1979) and William Sharpe (1976) but has recently been stated more directly by Jeremy Bulow (1982).[2] In essence, their views rest on a "legal" interpretation of pension liabilities; namely, since the firm can legally terminate its pension plan anytime, the firm's pension liabilities cannot exceed the present value of the *nominal* pension promises it would owe workers if the firm terminated the plan immediately. If workers believe the firm will exercise its legal right, it is intuitively apparent and easy to show that workers will deposit only that amount of compensation into the pension plan that they expect to receive upon immediate termination. Using this interpretation of pension plan liabilities, pensions have no influence on whether the worker leaves or remains in the firm. Moreover, it will be shown that if the pension contract is written in nominal terms, reflecting the legal view, workers are not bondholders in the firm: pension plans are excessively overfunded.

An alternative view is that the pension plan is an "implicit contract" between the worker and the firm (see, for example, Jack Treynor, 1977). An implicit contract theory implies that workers anticipating careers with a firm will consider the package of wage and pension benefits they expect to collect over their life cycle. It is shown that these expectations lead quite directly to several implications: that the pension contract is written in real terms; that firms intend to meet their pension promises; and that the firm will re-

*U.S. Department of Labor, S-4521, Washington, D.C. 20210. The views expressed in this paper are my own and do not necessarily represent the views of the Department of Labor. The paper has substantially benefited from the comments of Pauline Ippolito and discussions with Jeremy Bulow. I am also indebted to Robert Clark, Gary Fields, Alan Gustman, Olivia Mitchell, James Pesando, John Turner, Michael Ward, and participants at the Labor Relations Workshop at Cornell University and the Law and Economics Workshop at the University of Chicago for comments.

[1]See my 1985a book.

[2]The argument is also implicit in other related papers; see, for example, Martin Feldstein and Randall Morck (1982).

tain the plan intact and pay pension liabilities that are either implicitly or explicitly indexed to the final wage.

Using the implicit contract model, workers in pension plans have strong incentives to remain with the firm. Moreover, compared to a legal model, workers expect to "get more" from their pension plan in this model, and hence contribute more heavily to the pension early in their careers. This savings propensity has strong implications for the wage-service profile; in addition, pension liabilities are much larger than those predicted under a legal model. Indeed, under this interpretation, it turns out that pension plans are significantly underfunded: workers are long-term bondholders in the firm, suggesting that pensions signal a more complex worker-firm contract than typically considered, one which involves investment in the firm beyond human capital investment.

I. An Economic Model of Pension Liabilities

The first task is to derive a formal expression for the economic pension liability incurred by the firm, one which captures the essence of the legal and implicit contract theories. This derivation will provide the basis for drawing empirical implications from the models.

A. *The Basic Model*

I begin by making the assumptions that firms do not provide pensions to workers for free, and that workers will not sacrifice wages in excess of the true value of the pension. Workers pay firms an amount that is precisely equal to the present value of expected pension payments. They forgo a portion of their total compensation throughout their work lives in the firm in exchange for a pension at retirement. An economic pension liability arises when the firm accepts a worker's implicit pension contributions; the liability is absolved only when the firm returns these contributions, plus accumulated interest, when the worker retires. If these savings are not held in trust, the transaction is tantamount to an unsecured loan agreement.

If pension savings could be observed directly, the calculation of firm pension liabilities would be a trivial task. Unfortunately the pension savings flow is a hidden component of compensation. The trick is to infer pension savings rates from announced pension benefits.[3] That is, if workers' implicit pension contributions equal the present value of expected pension benefits, presumably it is possible to infer what pension savings rates *must be*, given workers' evaluations of announced benefits. Proceeding in this backwards fashion provides information about the worker's savings flow over the tenure cycle. The portrait of this flow will be radically different depending upon whether one believes in a legal or implicit contract theory. It is this distinction that is exploited empirically to determine which savings flow, that is, which theory, best describes reality.

To illustrate, suppose a firm has one worker who starts at the firm at age 0 and retires at age R. Thus age and service level can be denoted by a, $0 \leq a \leq R$. In reality, a worker typically vests at some age a^0, $0 < a^0 < R$. If the worker dies or quits the firm or if the firm fails or terminates the pension plan before he attains the age and service level a^0, he collects nothing; if either of these events occurs between the ages $a^0 < a < R$, he collects the nominal value of his pension accrual as of service level a, payable at age R.[4] Regardless of the theory of the pension contract, however, liabilities attributable to unvested workers are very small. Thus, I can

[3] It is theoretically possible to measure these savings rates by estimating the cash wages of pension-covered vs. uncovered workers over all ages. In practice, for a variety of reasons, this task has proven to be difficult and has not been successfully accomplished to date.

[4] If the firm fails early *and* does not maintain sufficient assets in the plan to cover at least its *nominal* liabilities, the expression becomes somewhat more complicated. Thus, for present purposes it is assumed implicitly that the firm maintains at least this level of funding. As shown below, most pension plans' funding status in fact do fall within the constraints of this assumption. In addition, since the passage of the Employee Retirement Income Security Act of 1974, pension-covered workers pay for insurance that guarantees receipt of nominal benefits in the event of plan termination.

simplify the discussion without materially affecting the results by considering only the case of vested workers.[5]

Now consider the pension formula. Suppose the worker's pension at age R is given in the form of a lump sum, and that the amount of the worker's pension is proportional to his years of service and the compensation level he is earning at the age he leaves, say age j.[6] Usually the pension formula incorporates the worker's *cash* wage at retirement. But it greatly simplifies the exposition without affecting the results if it is supposed the formula incorporates the worker's *gross*-of-pension-savings wage (i.e., cash wage plus the worker's implicit pension savings during that year); denote this compensation by W_j. If it is assumed that pensions are the only fringe benefits, W_j can be thought of as the individual's value of marginal product at age j. Thus, the pension formula can be represented by

$$PB = baW_j, \tag{1a}$$

where PB represents the lump sum pension benefit, payable at age R, and b is a constant reflecting the generosity of the pension plan.[7]

It is easy to characterize the worker's expected pension benefits at any given age; they depend on the pension formula and the probabilities that the worker will leave the firm early, or that the firm will terminate the plan early. If the worker is vested, the present value of the pension can be represented by

$$PV_a = \int_{j=a}^{R} baW_j f_{ja} e^{-i(R-a)}\, dj, \tag{1b}$$

where i is the nominal interest rate, j is the termination age; and f_{ja} is the conditional probability density of pension accruals stopping either because the individual leaves the firm or the firm terminates the pension at age j, given that the individual is currently age a.

The firm's pension liability when the worker is age a depends on how much the worker has saved in consideration of receiving this pension, which in turn depends on the worker's evaluation of the value of his promised benefits. The key to the worker-firm contract riddle is held by the density parameter f_{ja} in (1b). What is the worker's assessment that, at retirement age R, he will actually receive a pension indexed to his final wage W_R compared to some lesser pension indexed to some lower wage W_j, $j < R$?

Consider the legal theory first. This theory is built on the presumption that workers believe the firm will in fact terminate the pension immediately (or they believe they will either quit or be fired imminently).[8] In this case, the density function in (1b) becomes a point density at the current age $j = a$, and the present value of the pension perceived by a worker age a is

$$PV_a = baW_a e^{-i(R-a)}. \tag{2a}$$

Now consider the implicit contract model. The worker believes the firm will maintain the pension plan intact, even though the firm

[5] Examination of annual pension report data using techniques discussed below suggests that the distortion is in the range of 10 percent. The distortion is small in the implicit contract model partly because unvested workers have low service levels and partly because the turnover among unvested workers is relatively high. For example, the annual quit probability of an unvested male worker under age 30 in a pension firm is between 10 and 25 percent, depending upon the worker's particular age and tenure level (see James Schulz, 1980). In the legal model, the distortion is low because use of a high discount rate virtually evaporates benefits accrued by unvested workers who are usually far away from retirement age.

[6] In fact, defined benefit plans almost always pay benefits in the form of annuities; but the results are not affected by this simplification.

[7] Most pension plans have benefit formulas that are similar to this specification. Pensions are not always dependent on the final year's salary. Oftentimes pensions will be based on the last three or five years of tenure. See, for example, Bankers Trust (1980) and Urban Institute (1982).

[8] Or if workers are so myopic they believe their wage will never rise beyond W_a regardless of inflation, overall productivity growth or the firm's age/service wage profile.

could legally terminate the plan at any time.[9] To keep the expressions neat, assume that the empirically small probabilities that the vested worker will quit or the firm itself will fail are approximately zero. It will be shown below that this assumption has no important consequences for the main results.[10] Using these simplifications, the density function in the implicit contract model is again a point density, but this time at the retirement age $j=R$, not at the current age $j=a$. As a result, the present value of the pension in (1b) can be written as

$$PV_a = baW_R e^{-i(R-a)}. \tag{2b}$$

Note that the implicit contract expression in (2b) differs from legal expression in (2a) only because the expression in (2a) incorporates the *current* compensation level W_a, while the expression in (2b) incorporates the *anticipated* compensation level at retirement W_R. To compare the expressions more easily, it is convenient to rewrite (2b) in terms of current compensation at age a. For this purpose, assume that the worker expects his level of compensation to increase over his age at the rate g; these increases will reflect the firm's wage-tenure profile, overall productivity growth and inflation. In this case, the expression in (2b) becomes

$$PV_a = baW_a e^{(g-i)(R-a)}. \tag{2b$'$}$$

[9]The credibility of the promise need not rely solely on a belief that history will repeat itself. Workers and firms presumably understand the wage penalty that would be imposed on any firm that reneged on its pension promises.

[10] Empirically, the vested quit-plus-death rate is approximately 2 percent. The annual death rate for the typical pension participant (starts with the firm at age 38, retires at age 62) is .0075; the quit rate for this age and service group is approximately one percent (Schulz). Comparison of plan termination data from the Pension Benefit Guaranty Corporation and Department of Labor pension universe data show that the probability that a pension-covered worker will find himself in a firm that terminates over a 25-year period is only 3 percent. It is shown below that since the quit probability is small, the implications of the zero quit assumption remain essentially unchanged (see fn. 14).

The present value of benefits at age a can now be written succinctly to accommodate both these theories as

$$PV_a = baW_a e^{(\lambda g-i)(R-a)}, \tag{2c}$$

where λ is a parameter that can be set equal to zero or unity. In a legal interpretation of pensions $\lambda=0$; in an implicit contract model $\lambda=1$. This is the expression I set out to derive. It says that the worker's calculation of his pension wealth is a discounted value of the benefit formula based on current service and wage levels. The only difference between the legal and implicit contract theories is the value of the discount rate; given substantial inflation this difference can be very large.

In the legal theory, the worker is so pessimistic that he believes that either the firm's termination of the pension plan or his own departure from the firm is imminent. In this theory, the wage index g in the economic calculation in (2c) is irrelevant to the worker; hence, expected pension benefits are based on current wages discounted by the *nominal* interest rate. In an implicit contract theory, the worker believes the firm will keep the pension plan intact regardless of its legal right to terminate. In this theory, the wage index is a critical component in his economic calculation; hence expected pension benefits are based on current wages discounted by the *difference* between the nominal wage growth rate and the nominal interest rate.

B. *Implications of the Model*

Given the derivation of the expression in (2c), some empirical implications from the model can now be derived. First, implications for the theory of worker-firm attachments are considered. Second, implications for the shape of the (cash) wage-tenure profile are examined. The pension-mobility relation has already been estimated elsewhere. The pension-wage profile implications are tested below.

1. *Pensions and Mobility.* The implications of the legal and implicit contract theories for the worker's propensity to stay with the firm

should be apparent. The present value of the worker's cumulative implicit pension savings to age a are given by equation (2c). If a worker leaves the firm at age a, the firm freezes the pension benefit in nominal terms; that is, upon premature departure, the firm sets the wage growth parameter g equal to zero. The worker collects a pension at age R but its value is indexed to his termination wage at age a.

The capital loss to a vested worker from early departure from the firm is, therefore,

$$(3) \qquad CL = PV_a(\lambda=0)\left[e^{\lambda g(R-a)}-1\right].$$

It is apparent from (3) that in a legal theory of pensions ($\lambda=0$) the capital loss from early departure is zero ($CL(\lambda=0)=0$); that is, since the parameter λ is set to zero in this theory, setting g equal to zero upon premature departure is inconsequential. Put another way, in a legal model, the worker pays only the incremental value of his nominal pension from each additional year's work (he disregards positive values for g); thus, upon departure, he gets exactly what he paid for.

In an implicit contract model, the capital loss from departure can be very sizable. In this model, the worker pays for an indexed pension. But if he quits, he receives only a nominal pension. In terms of the expression in (3), the worker sets the parameter λ equal to unity in making his savings decision in the firm. Upon departure, setting g equal to zero frustrates this expectation. If a worker is required to implicitly contribute to the pension plan in expectation of receiving an indexed pension upon retirement, he has a powerful incentive not to leave the firm early. Consider a worker with 20 years of service and 10 years until retirement ($a=20$; $R-a=10$); also assume that the lump sum pension is worth 15 percent per year of service ($b=.15$) and the nominal interest rate equals the nominal wage growth rate which in turn equals 10 percent ($i=g=.10$). In this case, $CL=1.88W_a$: the capital loss from early departure equals almost two times his current annual wage.

Contributions in the literature tying reduced mobility to pensions implicitly accept the implicit contract theory (see for example, Olivia Mitchell, 1982; Bradley Schiller and R. D. Weiss, 1979): workers pay for real pensions but upon departure collect only nominal pensions. The legal model predicts no relationship between pensions and mobility: quitters receive precisely what they paid for. The empirical results showing a negative relation between pensions and mobility indirectly confirm the validity of the implicit contract model and reject the legal model. More direct empirical tests can be taken from the theories' predictions about wage-service profiles.

2. *Pensions and Wage-Service Profiles.* The implications of the two theories for the shape of the wage-tenure profile are only slightly less apparent. To make this discussion concrete, however, one additional assumption is made. Following Bulow, it is assumed that workers' cash wage plus pension savings is equal to the value of marginal product in *each* period:

$$(4) \qquad CW_a = W_a(1-S_a),$$

where CW_a is the cash wage at age and service level a, and S_a is the change in the present value of the pension (net of interest earnings) as a percent of total compensation at age a. If this relation does not hold—if the firm arbitrarily imposes wedges between the pension accrual rates and the actual savings flow—the predictable effects of the legal and implicit contracts for cash-wage profiles are lost. In this sense, it is noted that the implicit contract theory in this model does *not* require abandonment of the spot market assumption. The "spot" assumption in (3) will be tested with the legal and implicit models below.

Consider the cash-wage profile in equation (4). To translate the expression into a more useful form, the pension savings rate S_a must be solved explicitly. Given the assumptions that workers pay for their pension accruals as they earn them, the savings rate is derived by differentiating the present value pension amount shown in (2c) with respect to age, and subtracting interest on past contribu-

tions.[11] Performing this exercise, the implicit savings rate at age a (as a percentage of total compensation) is

$$(5)\quad S_a = [1+(1-\lambda)ga]be^{(\lambda g - i)(R-a)}.$$

This expression plus the spot market assumption in expression (4) provide a means to test the legal and implicit contract theories. It is apparent from (5) that in a legal theory ($\lambda = 0$), the pension savings rate increases rapidly at higher service/age levels in the firm: $S_a' > 0$ and $S_a'' > 0$. Under plausible conditions, the legal theory can imply drastic backloading of savings late in life. For example, suppose the nominal growth rate in wages and the nominal interest rate are each 10 percent ($g = i = .10$), the tenure level at retirement R is 30 years, and the pension benefit parameter b is .15.[12] In this case, the savings rate at starting age ($a = 0$) according to equation (5) is less than 1 percent; at age 30, the savings rate is 60 percent. Intuitively, the distrustful worker is a reluctant pension saver until he actually approaches retirement. But if this is so, the wage relation in (4) suggests that cash wages should become abnormally low at later tenure levels because pension savings consume progressively larger portions of gross compensation.

In contrast, the savings rate in the implicit contract model is not dependent solely on the nominal interest rate i, but rather on the *difference* between the wage growth rate and the interest rate, $i - g$. To pursue the full implications of this model, it is useful to acquire a plausible parameter value for this difference. It turns out that as long as the long-term interest rate and expected wage growth equally reflect inflation, the difference $i - g$ will be significantly less than zero: the real rate of return earned by pension funds is substantially lower than real wage growth, by approximately two percentage points.[13] If the model is generalized to allow early quitting, firm failure, and premature death, it can be shown that in the extreme, the sum of these probabilities would act exactly like higher real interest rates. If one thinks of the interest rate as including these additional influences, it still would not dominate the influence of real wage growth.[14] Thus, in terms of the current notation, it is generally true that real wage growth weakly dominates an enhanced interest rate: $g \geq i$.

Consider the plausible case where the nominal growth rate in wages equals the nominal interest rate, $g = i$. In this case, the implicit contract model predicts that the change in the savings rate with respect to service level is zero: $S_a' = 0$. When the worker and firm have an implicit long-term contract, the worker is willing to save through the pension more evenly over his life. In terms of the cash-wage expression in (4), the implicit contract theory says that pensions should affect cash wages proportionally over the tenure cycle. Pensions have no basic influences over the shapes of these profiles.

The wage-service profiles provide a clear test of the theories. Consider a comparison

[11] In particular, even if no additional savings takes place, just by waiting one additional period in the firm, the individual receives interest on past contributions equal to iPV_a.

[12] The average savings rate of covered workers through their pensions is approximately 15 percent. See my 1985b article.

[13] The real rate of return earned by pension funds over the longest period over which data is available (1963–82) has been approximately 1 percent (Alicia Munnel, 1983). Wage growth reflects a presumably positive wage service profile in the firm as well as overall productivity growth. Roger Gordon and Alan Blinder (1980) have estimated a wage growth rate over the life cycle of work with the firm (holding firm productivity constant) in the vicinity of 1 percent. Real wage growth in the United States over the past 30 years has been in the vicinity of 2 percent (*Economic Report of the President*, 1982). Thus, assuming that wages and interest similarly reflect an expected inflation factor, it follows that $g - i$ is approximately 2 percent.

[14] To show this, take the extreme assumption that departure from the firm early or firm failure will have the same result as early death: no pension will be paid (instead of a nominal pension). In this case, the sum (say q) of the quit rate, death rate, and firm failure rate will enter the model exactly like the interest rate. Since q is almost certainly less than 2 percent (see fn. 10), we still have $g \geq q + i$ (see fn. 13). Introducing this factor makes the equality a more likely approximation to reality, but does not change the essence of the qualitative result: discounting as such will not occur during the worker's tenure in the firm.

of cash-wage-service profiles for pension-covered and uncovered workers. Other things constant, if the legal theory is correct, these profiles should be flatter for pension-covered workers. If the implicit contract theory is correct, these profiles should be either no different or perhaps even somewhat steeper for pension-covered workers.

II. Tests of Pension Contract Theories Using Wage-Tenure Relations

A. *Specification of a Wage-Tenure Equation*

For empirical purposes, I choose a functional form for the wage function that is consistent with a positive but diminishing tenure effect, but that also leads to a simple representation of the contrary predictions of the legal and implicit contract theories. I wish to make specific allowances for the possibilities that pension savings rates increase (or decrease) over the tenure cycle and that the spot market assumption in (4) is invalid. A wage function that satisfies these criteria is

$$(6)\quad CW_T e^{\{f+[\kappa+\gamma \ln T]P+[\delta(B)+l(B)\ln T]\}} = kT^{\alpha+\beta X}X^c e^{dP}, \qquad T>0.$$

This specification is somewhat unconventional because it implies that the start wage is zero; while this could be a problem in theoretical models, it is not important for my purposes as long as the data is concentrated away from zero tenure levels.

The left-hand side of (6) represents total compensation for the worker at the tenure level T; the right-hand side represents the value of marginal product. The variable CW_T is the worker's cash wage; total compensation equals the cash wage, adjusted by the series of terms found in the exponential term. The parameter f reflects the nonpension fringe benefit as a percent of cash wage. The first bracketed term in the exponent that is interacted with a zero-one pension dummy P ($=1$ if there is pension coverage) reflects the pension-savings rate. The parameter κ is a constant. The parameter γ is critical to a test of the theories: it reflects the influence of tenure on the pension savings rate. In the legal model, the wedge between gross and net-of-pension-savings compensation grows with tenure: $\gamma >> 0$; in the implicit contract model (recall $g \geq i$), the parameter γ is either zero or slightly positive.

The second bracketed term in the exponent reflects a test of the spot market assumption. Total compensation may not equal value of marginal product in every period. For example, the firm could deliberately twist the cash wage profile so that workers get paid too little early and too much later (see Edward Lazear, 1979). What makes this theory testable, however, is that compensation over the worker's tenure must equal his value of marginal product. But while workers in any given firm tend to retire around the same age, they begin over a wide range of ages.[15] It is easy to show that in order to balance the books in a nonspot model, the parameters δ and l in (6) must be set differently depending upon the worker's beginning age in the firm, denoted by B.

The right-hand side of (6) is straightforward. The worker's marginal product is assumed to be related to his tenure level. The vector X is permitted to affect compensation directly and to influence the shape of the wage-service profile. The specification in (6) allows pensions to affect (or to be correlated with) productivity in a proportional way through the term: dP.[16]

[15] Work patterns in firms are clearly not homogeneous. For example, in a Department of Labor sample of 80 firms that reported at least 25 retirees during the period 1975–78, the standard error on age retired within the firm averaged 3.5 years. The standard error for service length in the firm averaged 13 years. Since there appears to be a fairly narrow window around retirement age, and since tenure levels have a wide variance, it is apparent that, in order for the firm to pay appropriate present value compensation to all workers, it must adjust workers' profiles according to the age started with the firm. Evidence from this same survey reported below contradicts this prediction (see Table 1).

[16] This term could also reflect the indirect impact of pensions on the rate of hours worked, an effect caused by lifetime tax effects affiliated with pensions. If the wage rate is related to the rate of hours worked, it can be shown that pension coverage can quite plausibly cause workers to follow (proportionally) higher wage profiles over the tenure cycle. See my 1985b article.

The most direct way of inferring the value of the parameter γ is to estimate the specification in (6) directly. Unfortunately, it is very difficult to estimate this equation using a cross section of workers because the selection on who accepts pensions will bias the estimate of γ (high-wage earners are more likely to accept pensions compared to low-wage earners). By exploiting individual wage history data, this problem can be circumvented. In particular, by comparing beginning and ending wages for each worker rather than comparing wages between pension-covered and uncovered workers, an estimate of the parameter γ can be made without encountering the nuisance selection problem present in a direct estimate specification. Writing the expression in (6) for the year of retirement and dividing by the same expression for the first year of work in the last job, and assuming $l(B)$ is a linear function, we have

$$(7)\quad \ln(CW_R/CW_0) = c_0 + c_1 \ln(T_R/T_0) + c_2 \ln(T_R/T_0)\cdot X + c_3 \ln(T_R/T_0)\cdot P + c_4 \ln(T_R/T_0)\cdot B + error.$$

The cash wages are adjusted by the nonagricultural wage index for the United States; hence the wage measures are purged of the influences of inflation and overall productivity growth. It is noted that in the transition from (6) to (7), several terms conveniently cancel, including the terms $\delta(B)$, X^c, and $\exp(P\,[d-\kappa])$; all remaining variables interact with relative tenure levels (except health problems at retirement which are specific to years late in the tenure cycle).[17]

In terms of the empirical model specified in equation (7), the legal theory of pension savings is supported if $c_3 = -\gamma < < 0$; in contrast, under plausible assumptions ($i = g$), the implicit contract theory is supported if $c_3 = -\gamma = 0$. If $i < g$, the implicit contract theory is also consistent with the result $c_3 > 0$. Finally, the spot assumption is supported if the coefficient on the beginning age variable is zero ($c_4 = 0$).

B. *Social Security Retirement Data*

To test these predictions, equation (7) was run using two data bases; the first is the Social Security Administration Newly Entitled Beneficiaries (*SNEB*) *Survey*. The advantage of this data base is that it includes information about pension-covered and uncovered workers. It does not include multiple observations per firm and thus no information is available to test the spot assumption. This shortcoming will be rectified (at a cost) using a second data base.

The *SNEB Survey* describes the personal and job characteristics of several thousand recently retired workers who applied for Social Security benefits in 1970. Years of service in their last main job are reported. Their earnings histories are also reported as far back as 1951. Thus, if the worker's tenure with the firm began prior to 1951, the beginning cash wage equals his earnings in 1951; his tenure variable is set equal to his tenure level in 1951. If workers reached the Social Security earnings maximum, their earnings are estimated based on an algorithm developed by Alan Fox (1982).[18] The vector of

[17]The empirical specification can be directly related back to the pension model used above. Using the expressions found in (4) and (5), the ratio of cash wages at retirement relative to the start age divided by the same ratio of total compensation is

$$\Omega = (CW_R/CW_0)/(W_R/W_0)$$
$$= [1 - b(1+(1-\lambda)gR)]/[1 - be^{(\lambda g - i)R}].$$

[18]Separate analyses of Fox's wage estimates against observable earnings data in the *Survey* reveal no bias in the estimation method. Actual ending (but not starting) wage earnings in the worker's last main job are reported in the *Survey*. Comparing these actual earnings to the Fox-estimated earnings for those facing the Social Security maximum earnings in their last year of full time work, the Fox-to-actual ratio turned out to be .965; thus, it is apparent that Fox's estimation method (based on quarters-of-coverage information in the Social Security files) comes on average remarkably close to predicting actual earnings. In any event, the regressions in Table 1 were run separately using only those observa-

other variables X in (7) includes the worker's occupation, industry, and personal characteristics.

The results of the regression are listed in the first column of Table 1. The estimates reveal a coefficient on the pension variable $PENSION(=P)$ that is slightly negative but insignificantly different from zero, suggesting that cash-wage profiles are unaffected by pension coverage; that is, implicit pension savings rates are approximately constant over the tenure of pension-covered workers. In short, the results support the implicit contract theory and contradict the legal theory.

C. *Department of Labor Pension Data*

These results can be verified using alternate data. In particular, the Department of Labor maintains a data base that reports (Social Security) wage histories for 1642 retirees from 61 defined benefit pension plans. While the data excludes firms with no pensions, the wage profile test can still be performed by substituting the generosity of the pension plan (*PENSION SIZE*) instead of a simple pension dummy variable. Pension generosity is measured by the parameter b used above (see equation (1a)) that is easily estimated for each plan.[19] From (5), it is apparent that in a legal theory, as the generosity of the plan is increased, the pension savings wedge that grows with tenure becomes even more exaggerated: $\partial^2 S_a/\partial a \partial b > 0$; in an implicit contract model (in the plausible case when $i=g$), the wage-service profile is either unaltered: $\partial^2 S_a/\partial a \partial b = 0$; or, if $g>i$, $\partial^2 S_a/\partial a \partial b < 0$. Thus, if the coefficient on *PENSION SIZE* is negative (and large), the legal theory is supported; if it is zero or positive, the implicit contract theory is supported.

This data base has several advantages over the Social Security data. For example, since all the individuals are covered by pensions, the selectivity problem on who seeks pension coverage is reduced. The data base also reports the union status of pension plan participants and the date of plan creation; thus, union influence on wage profiles can be measured, and further, all observations for which the starting year wage (CW_0) is observed is earlier than the date of plan creation can be excluded from the data.[20] Finally, because each of the 61 plans in the sample has an average of 27 retirees,[21] the influence of the average wage and the average starting age in the firm on wage profiles can be measured. The starting age of each worker relative to the average in the firm can be used as a basis to test of the spot equilibrium model.

The results of this regression are presented in the second column of Table 1. Again, the results contradict the legal theory of pensions. Far from being negative and large, the coefficient on *PENSION SIZE* is positive. The results are consistent with the implicit

tions that were not estimated by Fox. The qualitative results (not reported) remained essentially unchanged.

[19]Most large pension plan formulas resemble the specification in equation (1a); that is, the pension benefit is log-linear in service level and wage at retirement. Benefits are usually actuarially reduced for "early" retirement relative to normal retirement age and joint and survivor election, and increased for "late" retirement. Normally, benefits are higher in firms that have later than normal ages of retirement. As such, for estimation purposes, pension plan formulas can be represented by the equation $\ln PENSAMT = a_1 \ln WAGE + a_2 \ln SERVICE + a_3\, JOINT\, AND\, SURVIVOR\, ELECTION + a_4\ AGE\ RETIREMENT + a_5\ EARLY + a_5\ LATE + a_6\ YEAR\ RETIREMENT + b_1\ PLAN\ 1 + b_2\ PLAN\ 2 + \cdots + b_{61}\ PLAN\ 61$, where *PENSAMT* is the retiree's annual pension as of 1978. *EARLY* and *LATE* refer to retirement before or after the plan's "normal" retirement age, thereby signaling an actuarial adjustment. Accounting for early and late retirement, *AGE RETIREMENT* refers to the normal age of retirement in the plan. The variable *WAGE* is the worker's final wage in the firm, and *SERVICE* is the individual's service in the firm at retirement. Finally *PLAN i* denotes a dummy variable equal to unity for all observations in the ith plan, zero otherwise. Estimating this equation using 1642 retirees in 61 plans, the 61 b-parameters were estimated.

[20]In the Social Security Administration data file, it was assumed that retirees were covered in the year their first wage was reported. Statistically, two-thirds of all pension plan participants were in plans established prior to 1951. Thus, some noise in the pension variable is present in that data.

[21]To prevent some large plans from dominating the results, no more than 100 retirees were selected randomly from the large plans.

TABLE 1—THE EFFECT OF PENSION COVERAGE AND PENSION AMOUNTS ON WAGE PROFILES

Independent Variables	Social Security Administration Retirement Data (1)	Department of Labor Pension Data (2)
Log of Tenure Ratio: $\ln(T_R/T_0)$	.030 (1.06)	$-.075$ (1.11)
Health Problems at Retirement	$-.020$ (1.53)	–
All Remaining Variables are Interacted with $\ln(T_R/T_0)$		
PENSION	$-.004$ (.40)	–
PENSION SIZE[a]	–	.129 (2.93)
College Graduate	.007 (.41)	–
High School Graduate	$-.015$ (1.06)	–
Female	.060 (4.82)	.018 (1.56)
Black	$-.001$ (.02)	.054 (3.82)
UNION[b]	–	.002 (.67)
UNION·PENSION SIZE	–	$-.029$ (.64)
Average (Annual) Wage in Firm[c]	–	.0086 (4.52)
Average Starting Age in Firm	–	.002 (1.81)
Worker's Starting Age Minus Average Starting Age in Firm	–	$-.0006$ (.94)
Industry Dummy Variables[d]	X	X
Occupation Dummy Variables[d]	X	X
Firm-Size Dummy Variables[e]		X
Year of Retirement Dummy Variables[f]		X
Observations	3413	1642
R^2	.054	.106

Notes: Dependent Variable: $\ln(CW_R/CW_0)$ Data for the first regression are from the *SNEB Survey* 1970. Data for the second regression are from a Department of Labor survey that includes Social Security Administration earnings histories of workers from 61 pension plans who retired during the period 1967–77; (absolute) t-statistics are shown in parentheses.

[a]*PENSION SIZE* is found by estimating the parameter b in the pension formula for each of 61 plans represented in the sample. See equation (1a) and fn. 19.

[b]*UNION* equals one if the pension plan is collectively bargained, zero otherwise; 37 of 61 plans were characterized by *UNION* = 1.

[c]Sample mean \$18.4 thousand. The average wage in the firm is taken as the average of (real) final wages of the sample of retirees in each plan in the sample.

[d]Eight industry dummy variables and eight occupation dummy variables were included.

[e]Seven firm-size dummy variables were included.

[f]Nine year-of-retirement dummy variables are included.

contract theory if $g > i$. Assuming the validity of the implicit contract theory, it is straightforward to determine that the results suggest that the exponent in (2c), namely $(i - g)$, equals $-.005$.[22] It is also interesting to note from the results that, while wage profiles are affected by the average wage level and the average start age in the firm, they are not sensitive to the relative start age of workers in each firm, providing some evidence in support of the spot model.

D. *Implications of Results*

The results using either data base are decidedly inconsistent with the legal theory of pensions that predicts a negative and large value for the pension coefficient. Instead, the results support the implicit contract theory in the special case when the interest rate is approximately equal to, or slightly lower than, the expected growth in nominal wages.

These results have important implications for the nature of the labor contract. First, the results offer strong support for theories that have pensions acting as a deterrent to early quitting or shirking (Lazear, 1982). If workers save in the firm in anticipation of receiving wage-indexed pensions upon retirement, premature voluntary or involuntary departure from the firm will trigger a sizable capital loss (see (3)).

Second, if workers save as if they expect a real pension upon retirement, it follows that, as a whole, workers must be long-term bondholders in the firm. That is, *reported* pension liabilities are usually calculated on the basis of the legal theory (pension promises are discounted at nominal interest rates). On this basis, reported liabilities are usually lower than pension assets: pensions are typically reported to be overfunded. But if the implicit contract theory is correct, reported liabilities significantly understate the true liabilities facing the firm. Using discount rates that are consistent with the implicit contract theory, it has been shown that pension plans are significantly underfunded (by approximately 30 percent during the post-World War II period);[23] this despite demonstrable tax advantages afforded full funding (Fisher Black, 1980; Irwin Tepper, 1981). In short, quite apart from human capital issues, the results imply that workers typically have a direct financial stake in the firm, one that takes the form of an unsecured long-term bond.

Third, and finally, a corollary to the worker's bondholding status in the firm is that the firm's stock value is reduced by the magnitude of the bond. That is, stockholders will rationally reduce the firm's stock price to reflect true economic pension obligations (Martin Feldstein and Randall Morck). Since underfunded economic pension liabilities represent approximately 25 percent of the firm's equity value,[24] it is apparent that the implicit contract theory has rather dramatic consequences for stockmarket value. Put another way, stock traders who have information about firms' true economic liabilities can make large amounts of money by trading with investors who fail to "see through" reported pension liabilities.

III. Concluding Remarks

In this paper, an economic model of pension liabilities was developed. The competing legal and implicit contract theories of pensions were derived from the model and empirical tests to distinguish the theories were developed. The empirical tests strongly support the implicit contract theory and contradict the legal theory. The firm appears to have an implicit contract with its workers that it will pay workers a real pension upon their retirement from the firm. As such,

[22] The coefficient on *PENSION SIZE* is an estimate of $\Omega'(b)$ (fn. 17). Calculating $\Omega'(b)$ from the expression in fn. 17 and setting $\lambda = 1$ (which assumes the validity of the implicit contract model), the implied estimated discount rate is determined.

[23] The derivation of true economic liabilities based on the implicit contract theory, together with funding distributions which are representative of the post-World War II period can be found in my 1985c article.

[24] This statistic is based on comparison of equity value and true economic underfunded pension liabilities for a random sample of firms in the Compustat data files.

workers begin to make sizable deposits into the pension plan early in the tenure cycle. The results support the notion that pensions create strong worker-firm attachments: premature departure from the firm imposes large capital losses upon workers. The same results, however, also lead to the conclusion that pension plans are significantly underfunded; economic funding ratios in the United States have persisted in the range of 70 percent throughout the post-World War II period. As such, the results confirm the authenticity of a puzzle: given the tax incentives to fully fund a tax-free savings account, why are pension plans underfunded?

By underfunding its pension plan, the firm makes its employees long-term unsecured bondholders in the firm. Such contracts could be optimal if workers have better information about the capabilities of the firm compared to outside lenders. But there is no reason to believe that workers better understand the financial viability of the firm than outside lenders who specialize in such evaluations. Moreover, to the extent that such employee loans to the firm represent substantial portions of their wealth, they would presumably require a higher risk premium than diversified outside lenders. Why would the firm presumably pay such a risk premium to engage in long-term lending arrangements with its employees rather than outside lenders?

A natural explanation for the contract lies in the control over the firm held by employees that is unique to them as a group. As a group, workers can substantially affect productivity and wage levels in the firm; hence, they can influence the financial viability of the firm over the long run. Outside lenders presumably cannot set conditions that would preclude adverse productivity actions by employees. The firm can arrange its pension rules to discourage poor performance by employees individually (see Lazear, 1982; my 1985a study). But when workers collude through a union, productivity and wages are more difficult to control.

The idea that it is efficient for the firm to make unionized workers bondholders in the firm is pursued more fully elsewhere (see my 1985d article). While those results are indeed consistent with this hypothesis, it is premature to say whether this idea or others will lead to a solution of the underfunding riddle. But whether it can be successfully solved or not, the evidence presented here strongly suggests that true economic pension liabilities are significantly higher than reported liabilities found in most publicly disclosed statements, sufficiently higher, in fact, to provoke a reconsideration of the fundamental nature of the worker-firm relationship.

REFERENCES

Barnow, Burt S. and Ehrenberg, Ronald G., "The Costs of Defined Benefit Pension Plans and Firm Adjustments," *Quarterly Journal of Economics*, November 1979, *93*, 523–40.

Black, Fisher, "The Tax Consequences of Long Run Pension Policy," *Financial Analysts Journal*, July-August 1980, *36*, 3–10.

Bulow, Jeremy, "What Are Corporate Pension Liabilities?," *Quarterly Journal of Economics*, August 1982, *96*, 435–52.

Feldstein, Martin and Morck, Randall, "Pension Funding Decisions, Interest Rate Assumptions and Share Prices," Working Paper No. 938, National Bureau of Economic Research, July 1982.

Fox, Alan, "Earnings, Replacement Rates and Total Income: Findings From the Retirement History Survey," *Social Security Bulletin*, October 1982, *45*, 3–23.

Gordon, Roger and Blinder, Alan S., "Market Wages, Reservation Wages, and Retirement Decisions," *Journal of Public Economics*, October 1980, *14*, 431–42.

Ippolito, Richard A., (1985a) *Pensions, Economics and Public Policy*, (Pension Research Council Monograph) Homewood: Dow Jones-Irwin, 1985 forthcoming.

______, (1985b) "Income Tax Policy and Lifetime Labor Supply," *Journal of Public Economics*, April 1985, *26*, 327–47.

______, (1985c) "The Economic Burden of Corporate Pension Liabilities," *Financial Analysts Journal*, 1985, *41*, forthcoming.

______, (1985d) "The Economic Function of Underfunded Pension Plans," *Journal of Law and Economics*, October 1985, *28*.

Lazear, Edward P., "Pensions as Severance Pay," Working Paper No. 854, National Bureau of Economic Research, February 1982.

______, "Why is there Mandatory Retirement?," *Journal of Political Economy*, December 1979, *87*, 1261–84.

Mitchell, Olivia S., "Fringe Benefits and Labor Mobility," *Journal of Human Resources*, Spring 1982, *17*, 286–98.

Munnell, Alicia H., "Who Should Manage the Assets of Collectively Bargained Pension Plans?," *New England Economic Review*, July/August 1983, 18–30.

Schiller, Bradley R. and Weiss, R. D., "The Impact of Private Pension Plans on Firm Attachment," *Review of Economics and Statistics*, August 1979, *61*, 369–80.

Schulz, James H., "Private Pension Policy Simulations," Report submitted to the U.S. Department of Labor, 1980.

Sharpe, William F., "Corporate Pension Fund Policy," *Journal of Financial Economics*, June 1976, *3*, 183–93.

Tepper, Irwin, "Taxation and Corporate Pension Policy," *Journal of Finance*, March 1981, *36*, 1–14.

Treynor, Jack L., "The Principles of Corporate Pension Finance," *Journal of Finance*, May 1977, *32*, 627–38.

Bankers Trust, *Corporate Pension Plan Study*, 1980.

Urban Institute, "Financial Retirement Incentives in Private Pension Plans," Report submitted to the U.S. Department of Labor, Washington, 1982.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington, 1982.

Rationing without Government: The West Coast Gas Famine of 1920

By ALAN L. OLMSTEAD AND PAUL RHODE*

> Why is it that for a century and more before 1971, there were no energy crises, no gasoline shortages, no problems about fuel oil—except during World War II? There is an energy crisis, there are gasoline lines, for one reason and one reason only. Because the government has decreed that there shall be.
> *Milton Friedman* [1979, p. 70]

The lesson is simple—without government bungling there would never be "shortages" or gas lines because prices would freely rise to clear the market. Many economists, including Friedman, have popularized this view, but there is a problem. In the spring and summer of 1920 a serious gasoline famine crippled the entire West Coast, shutting down businesses and threatening vital services. Motorists endured hour-long lines to receive 2-gallon rations, and, in many localities, fuel was unavailable for as long as a week at a time. This general situation continued for over four months, which is about as long as the 1973 and 1979 shortages lasted.

The troubling aspect of the 1920 crisis is that despite the outward indications of government intervention—priority users, two-tier pricing, ration cards, and quality deterioration—there were no government price controls or rationing programs. During one of the most rapid inflations in American history, western oil marketers voluntarily suppressed price advances and, instead, created and administered a complex allocation scheme. Thus, the beliefs that there were no energy shortages in the United States before the 1970's and that large-scale rationing requires government price controls are clearly wrong. Two intriguing questions arise: why did private entrepreneurs choose nonprice rationing, and, given that some businessmen opted to ration, why did the market allow them to succeed?

To shed light on these issues we will analyze the extent of the shortage, the nature of the rationing program, and the structure of the petroleum industry. We argue that regional isolation, industry concentration, and the vertical integration of the larger firms made rationing possible. In the absence of laws requiring rationing or setting prices, we focus on the hypothesis that the oil companies held prices down because they were afraid of hostile government actions.

I. An Overview of the Period

In late 1919 and early 1920, the petroleum market tightened nationwide as oil demand, fired by the postwar business expansion, soared. Domestic consumption jumped 25 percent in 1920, outstripping record domestic production and driving the real price of crude to levels unmatched over the next fifty years. On the West Coast, a separate oil market dominated by Standard Oil of California (SOCal), crude's real price doubled during 1920. As in the 1970's, government and industry experts predicted an imminent decline in U.S. oil production, and the ex-

*Director, Agricultural History Center and Professor of Economics, University of California, Davis, CA 95616, and graduate student in economics, Stanford University, Stanford, CA 94305, respectively. We have benefited from discussions and correspondence with M. Adler, J. Bender, R. Castanias, A. D. Chandler, Jr., V. Goldberg, P. Lindert, D. McCloskey, D. North, J. Pratt, K. Sokoloff, S. Sheffrin, J. Shideler, G. Stigler, G. Walton, and G. White. We also acknowledge the comments of seminar participants at UC Berkeley, Stanford University, UCLA, UC Santa Cruz, American University, and the University of Chicago. Finally, we have gained valuable insights from John Abel of the Union Oil Company and R. G. Follis, D. B. Willmer, and W. D. Hermann of the Standard Oil Company.

haustion of domestic reserves within a generation.[1]

In early April 1920, SOCal began urging motorists to "buy as little gasoline as possible," and soon thereafter gasoline lines, empty stations, stranded autos, and idled tractors became a common sight in rural areas.[2] These initial spot shortages, due in part to a wildcat railroad strike, grew quickly into a severe gasoline famine, engulfing the entire West Coast for the coming four months. As the situation worsened in late May, the major oil companies jointly forged a rationing program for all of California, designed to cut consumption by 15 percent. Their intent was to maintain supplies at historic levels to "essential users" and force "nonessential" consumers to bear the entire shortfall. Farmers, doctors, police, and businesses dealing in perishables would receive special priority while personal automobiles would be sold no more than 5 gallons at a time if fuel was available. Since "pleasure cars" consumed about 30 percent of the gasoline sold, the rationing program cut their allotment in half. Neither federal nor California state officials participated in creating or policing the rationing program. In Oregon and Washington, far more restrictive plans were imposed.[3]

Company sales representatives and special community boards usually decided allocation details given these general guidelines. Local officials often imposed Draconian measures. For instance, authorities in Wenatchee, Washington, grounded almost 100 automobiles after spotting them late one Saturday night "on pleasure bent." Police ordered distributors not to serve these "joyriders" for the duration of the shortage. Seattle police began arresting drivers who left their parked vehicles idling. Everywhere, individuals caught hoarding or using their privileged status to earn black market profits were cut off and even prosecuted. Many localities issued ration cards, and most major cities seriously considered it. In early June, oilmen encouraged Los Angeles officials to implement card rationing, but civic leaders instead enacted a program that set aside special stations for key segments of the business community. Hollywood, for example, had one station exclusively for the movie industry.[4]

Gas station owners were hassled by irate customers, traffic jams, and erratic work schedules. On July 2 in Oakland, California, 150 cars queued at one station impeding traffic over a four-block area. On July 16, Standard's stations in Long Beach relaxed restrictions so that after 2:00 P.M., all cars (as opposed to only commercial vehicles) could buy up to 2 gallons. This caused such a jam that "it was necessary for the assistant special agent to spend two hours in keeping the traffic from blocking the streets" (SOCal "Daily Reports, July 1920"). Drivers in the Pacific Northwest endured 2-gallon rations for several months, and even with these limits stations often closed by 9:00 A.M. In San Francisco, gunplay erupted in a dispute over ration entitlements.[5] Clearly, the decision to fix prices caused enormous headaches for oil company personnel from executives down to station attendants.

Holding prices down did not prevent public unrest. The press and many business groups declared open season on the oil companies, claiming that the shortage was contrived. Rumors circulated accusing the leading firms of hiding oil in remote storage facilities, exporting massive quantities of gasoline, and preparing for enormous price hikes. Farmers and automobile dealers campaigned to make gasoline a public utility and to end exports, even to neighboring states. "Beggar thy Neighbor" policies gained adherents. The *Los Angeles Examiner* (July 14,

[1]U.S. Geological Survey (*USGS*) *Mineral Resources ... 1921* (1924, pp. 253–54); ... *1922* (1925, pp. 261–63).

[2]For example see *S.F. Chronicle*, April 10, 1920; *Sacramento Union*, April 18, 1920 and May 17, 1920.

[3]For example see *Standard Oil Bulletin*, July 1920, p. 7; *Sacramento Union*, May 23, 26, 1920; *L.A. Times*, May 26, 1920.

[4]*Seattle Daily News*, June 4, 7, 1920; *L.A. Times*, June 2, 13, 1920.

[5]*S.F. Examiner*, July 3, 1920; *S.F. Chronicle*, July 8, 1920.

1920) called on SOCal to abrogate its contracts to supply the Hawaiian Islands, and in Turlock, California, the Board of Trade demanded that the oil companies cease deliveries to local Japanese farmers. Independent station owners and garage operators complained bitterly that the majors were attempting to force them out of business.[6]

On July 9, the *Los Angeles Times* reported the unbelievable: John D. Rockefeller, Jr., had to cut short his southern California vacation because "he cannot buy gas." He was not alone. In early July, cities from Seattle to San Diego came to a near standstill. Even in oil-rich Los Angeles, boulevards normally choked by traffic were empty. The sole consolation was that the traffic accident rate had plummeted to one-tenth of normal. In late July, the situation took a dramatic turn. Union Oil and a number of small independents began importing large quantities of midcontinent gasoline directly into Los Angeles, using the high cost of these imports to justify sharp price increases. The progress of the special relief trains became front page news as the end of the famine finally came within sight.[7] But the lines and rationing did not disappear completely until the end of August 1920 after the flood of imports, price increases, and a sharp midyear downturn in economic activity brought a return to normalcy.

II. Consumption, Production, and Prices

A vigorous postwar business expansion led by a virtual explosion in automobile sales pushed gasoline consumption to unprecedented levels. During the first six months of 1920, SOCal sold 70 percent more gasoline in its western market than it had sold during a similar period in 1918 without significantly changing its market share. In line with national trends, the stock of passenger cars and trucks on the Pacific Coast jumped from about 620,000 vehicles on December 31, 1918, to over 906,000 on December 31, 1920. By 1920, there was one auto for every six Californians. The number of tractors almost doubled in the eighteen months preceding June 1920. In addition, a severe regional drought in 1919–20 shocked the demand for gasoline outward. The dry weather encouraged winter motoring far above the normal rates, and caused a drastic shortage of surface water and hydroelectric power, forcing farmers to shift to gasoline pumps. The oil companies repeatedly asserted that this prevented them from replenishing their stocks over the winter. It also created a crisis atmosphere that undoubtedly contributed to the decision to ration fuel.[8]

A scarcity of crude prevented gasoline production from keeping pace with soaring consumption. In 1919 and 1920, crude output in California increased by less than 3 percent a year, one-fourth the national growth rate. The East Coast also had cheap access to a growing flow of Mexican crude.[9] Figure 1 shows quarterly production and refinery stocks of gasoline. In contrast to the general upward trend of the period, California production in 1919 actually fell nearly 4 percent. Gasoline output remained constant in early 1920 before rapidly rising in midsummer as several new fields opened. Refinery stocks plummeted from a level of about two months' production in early 1918 to less than two weeks' output in early 1920. A headlong fall in SOCal's inventories accounted for 95 percent of the drop in California refinery stocks during 1919. By contrast, national production steadily increased and stocks followed the normal cyclical pattern, rising in the winter and falling in the summer. The national reserve ratio never approached levels as low as those prevailing in California in 1920. To supply their western customers, both SOCal and

[6] SOCal "Daily Reports, July 1920"; *L.A. Times*, June 7, 1920.

[7] Earl Welty and Frank Taylor (1958, pp. 135–36); *Mining and Oil Bulletin* (Los Angeles), August 1920, 530–32.

[8] *Standard Oil Bulletin*, July 1920, p. 6; [Bureau of Foreign and Domestic Commerce], U.S. Department of Commerce, *Commercial Survey of the Pacific Southwest* (Washington: USGPO, 1930, p. 358).

[9] *USGS, Mineral Resources... 1921* (pp. 260–61; 309).

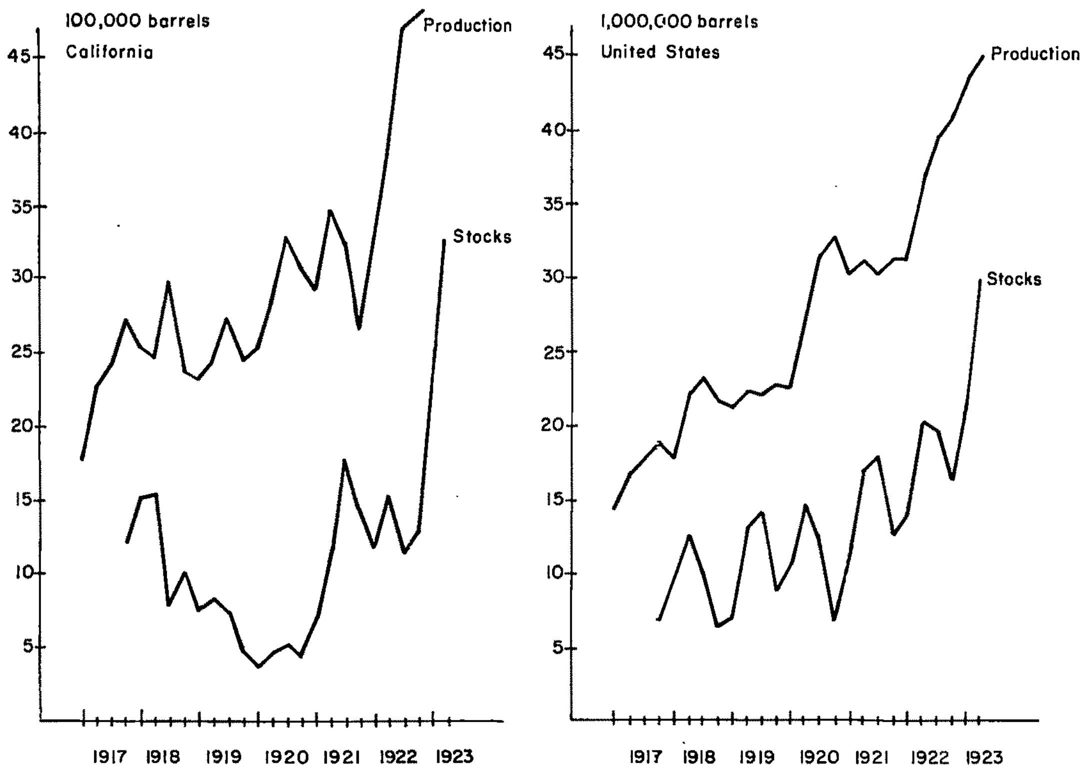


FIGURE 1. GASOLINE PRODUCTION STOCKS

Source: G. R. Hopkins (1927, pp. 50–53; 92–95).

Note: Gasoline Production and Stocks include other naphthas such as engine distillate.

Union restructured their normal marketing patterns by restricting foreign exports and altering their product mix in favor of gasoline. In addition, they imported unprecedented quantities of fuel from across the Rockies. For example, SOCal normally supplied its entire western market from California fields, but, in late 1919 and 1920, more than one-third of its gasoline was imported. An inability to obtain tank cars prevented SOCal from bringing in substantially more fuel.[10]

The most striking aspect of the crisis is the absence of price increases during a period of rapid inflation. Figure 2 graphs nominal retail prices for gasoline from 1917 to 1921 in San Francisco and in representative eastern points. Between August 1918 and July 1920, San Francisco prices advanced only twice. On June 10, 1919, SOCal raised its price one cent to 21.5 cents per gallon, and on March 17, 1920, it raised its price to 23.5 cents. In both instances, other marketers followed immediately. During the famine, despite persistent accounts of isolated stations selling

[10] In the spring and summer of 1920, SOCal's Traffic Manager, S. G. Casad, searched the country for spare cars taking "every tank car I could secure, regardless of the price...." (letter to H. M. Storey, Vice-President, SOCal, September 17, 1920, Box 120808; Sales Department, "Annual Reports, 1891–1922," p. 31, typed ms., Box 120762, SOCal Archives; Union Oil had similar problems (see their "Report of the Manufacturing Department, 1920," Appendix B, Refinery Operations, p. 16, typed ms.).

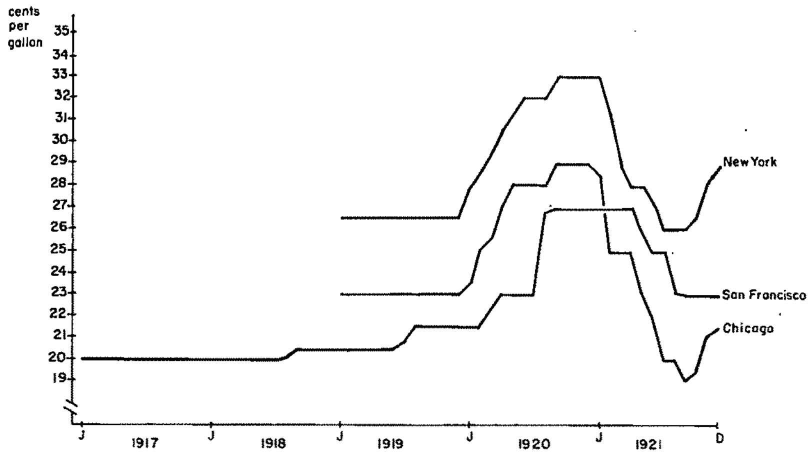


FIGURE 2. NOMINAL RETAIL PRICES FOR GASOLINE IN SAN FRANCISCO, NEW YORK, AND CHICAGO

Source: Compiled from *National Petroleum News*, 1916–1921.

gasoline for as much as a dollar a gallon, the price generally held firm. SOCal aggressively enforced a uniform policy among its retailers, cutting off fuel to those who overcharged or adulterated the product.[11] Not until mid-July did a few small refiner-marketers in Los Angeles finally break with SOCal. Some increased their prices by 7 cents, but the gas lines at their pumps remained. On July 24, Shell Oil followed independents, and over the next ten days, Union Oil inched its price up to 27 cents. SOCal waited until August 6 to match Union's advances. This was the first break with SOCal's price leadership in the gasoline market since 1915.[12]

Even these price increases confound the logic of microeconomics. Union Oil raised its price 3.5 cents only after it had embarked on a massive campaign to import gasoline from Oklahoma and Texas, where the retail price was a full 7 cents above that prevailing in Los Angeles. Standard followed a similar policy, charging an "average" price for the "mix" of local supplies and expensive imports. The private sector even invented two-tiered pricing systems. As one example, in the San Joaquin Valley, a motorist could purchase 3 gallons at 24.5 cents up to 9:00 A.M., and after that an unlimited amount at 35 cents a gallon. "The first was the local product at the old price, the last the imported fuel at the new figure."[13]

The rigidity of western gas prices is in sharp contrast to national trends. Between June 1919 and June 1920, the national cost-of-living index jumped 23.5 percent and then plummeted after June 1920. When SOCal was formulating its rationing program, her eastern sisters showed little reluctance to fetch what the market would bear. The absence of price increases on the West Coast is more remarkable given that western markets were far tighter than those in the East. Fur-

[11] *L.A. Times*, July 23, 1920. As an example: "a dealer at Caruthers recently raised his price... from \$.25 to \$.30 per gallon, but was informed that no further deliveries would be made to him unless the market price was adhered to. This he agreed to" SOCal "Daily Reports, July 1920."

[12] FTC (1922, Part II, pp. 65–70; 256–57).

[13] *Sacramento Bee*, August 4, 1920. Similar systems with price differentials up to 30 cents a gallon between the "old" and "new" gas emerged at several locations in the Pacific Northwest (SOCal "Daily Reports, July 1920").

TABLE 1—WEST COAST PETROLEUM INDUSTRY STRUCTURE
(PERCENTAGE SHARES OF LEADING FIRMS)

	Ownership of Proven Oil Acreage March 1, 1920 (1)	California Crude Oil Production 1919 (2)	California Refinery Output Jan.–June 1919 (3)	West Coast Gasoline Sales 1919[a] (4)
SOCal	10.9	25.5	47.6	55
Union Oil	10.0	8.6	16.5	20
Associated Oil-Southern Pacific	30.6	18.2	9.2	11
Shell Oil	3.1	6.6	10.1	4
Top 4 Firms	54.6	58.6	83.6	90

Sources: Col. 1: FTC, *Pacific Coast Petroleum Industry*; Part I: *Production, Ownership, and Profits* (1921, p. 45) (hereafter cited FTC, Pt. I): Col. 2: FTC, Pt. I (p. 64); Col. 3: FTC, Pt. I (p. 178); Col. 4: FTC, *Pacific Coast Petroleum Industry*; Part II: *Prices and Competitive Conditions* (1922, p. 58) and FTC, "Companies' Replies to the Dyer Resolution Questionnaire," Record Group 122, Boxes 3317-45.

[a] The shares of the domestic market can only be roughly approximated because no precise estimates of total sales on the West Coast are available for this period. Although the existing data on small marketing companies are incomplete, and do not satisfactorily distinguish between export and domestic sales, we can reasonably approximate their contribution. The figures presented for the four largest firms accurately reflect their relative sizes but probably understate their shares in total domestic sales.

thermore, California gasoline prices fell out of line with local crude prices. While the real price of light crude doubled in 1920, the real price of gasoline climbed only 20 percent. In fact, in the twelve months before lines appeared, the real price of gasoline fell almost 10 percent.

III. Market Structure

Market structure and institutions restraining competition helped make rationing possible. SOCal was, by far, the leading firm in the West. In this separate market, encompassing Arizona, California, Nevada, Oregon, and Washington, other Standard Oil companies refused to compete. Powerful geographic forces insulated the region as transportation costs between the midcontinent fields and the coast typically exceeded normal price differentials. Table 1 offers a view of the extent of the concentration and vertical integration of the West Coast petroleum industry. In 1919 four firms (Standard, Union, Associated, and Shell) produced roughly 59 percent of the crude oil and sold about 90 percent of the gasoline. SOCal headed the list, selling over one-half of the region's gasoline. Relative to national standards these firms were highly integrated, controlling much of their business from ownership of producing lands, to refining and transport, to rapidly growing networks of service stations.[14] In 1920, marketers exerted far more control over retail pricing than is possible today. As an example, Standard Oil's executives dictated prices for each locality, which all distributors, including independent garages, charged or risked being suspended.

Between 1915 and 1920, the West Coast petroleum market closely approximated the textbook price-leadership model. Since 1915, SOCal had initiated almost every gasoline price change. Private institutional arrangements helped maintain price stability. Galvanized by intense gas wars in 1915, most

[14] Gerald White (1962, pp. 486–519, 541–63); FTC (1921, Part I, pp. 42–65; 202–07, and 1922, Part II, pp. 30–74).

independents joined the Independent Petroleum Marketers Association, and formally agreed not to undercut Standard Oil's retail gasoline price. Such understandings were strengthened by the almost universal practice of naming Standard Oil's posted price for crude, as of the date of future delivery, as the contract price.[15] Even W. L. Stewart, president of Union Oil, found adhering to Standard Oil's prices "natural and unavoidable" (Gerald White, p. 497).

The industry's concentration, the centralized control over retail pricing, and the tradition of abiding by SOCal's prices provided an environment in which private rationing might succeed. Vertical integration and high profit levels allowed wide room for discretionary actions. SOCal's historic role as price setter probably created some initial bureaucratic inertia as other companies waited for SOCal to take the lead. While prices languished below market-clearing levels, the region's isolation and its preexisting distribution system prevented the rapid entry of outside producers.

IV. Did Government Pressure Firms to Ration?

Why did the oil companies suppress prices in the first place? Explanations suggested by conventional economic thought and by observers of the period fail to provide a convincing answer. One set of hypotheses holds that nonprice rationing was consistent with short-run profit-maximizing behavior for at least some firms. In 1920, many observers accused SOCal of suppressing prices to squeeze out competitors. But these charges make little sense, because nothing prevented small firms from raising prices and because the wide margin between SOCal's price and what consumers were willing to pay created profitable opportunities for entry. Furthermore, SOCal continued its pricing policy long after it was clear that 1920 would show record profits for almost everyone in the industry. In addition, internal company documents show that SOCal was in general exceptionally fair in maintaining deliveries to independent distributors even when its own stations had short supplies.[16]

Another class of hypotheses involves the possibility of adverse and costly government actions. Government price controls ended in late 1918, and wartime authorities never officially restricted gasoline use on the West Coast. In the absence of explicit laws, the oil companies may have feared hostile responses from independent bureaucrats or from special interest groups, which might have led to unfavorable government action.[17] This variant of the "government" hypothesis represents a major methodological retreat because it is probably impossible to test. While fear surely can be a potent economic force, it is inherently subjective. By assigning agents with the appropriate expectations or conjectures, one can explain almost any action as a consequence of rational calculation without gaining much understanding of it. Why would government threats work in this instance and not in others, when robust market forces make such threats and even laws ineffective? To assess the fear-of-government hypothesis, we will examine the general trend in industry-government relations, and ask how other firms behaved when faced with similar circumstances.

The oil industry emerged from the war with a lingering spirit of cooperation and with vastly improved relations with the federal government. Most wartime controls and industry boards were abolished shortly after the armistice (Gerald Nash, 1968, pp. 38–39). As direct federal involvement ended, the threat of antitrust prosecution also diminished significantly. On March 1, 1920, the Supreme Court issued its landmark "rule of reason" decision in the U.S. Steel case.

[15] FTC (1922, Part II, pp. 5–13, 30–31, 75–126, 135).

[16] For example see SOCal "Daily Reports, July 1920," Box 120773, SOCal Archives: letter, H. D. Collier, to all district sales managers, May 19, 1920; Memo from H. D. Collier, "Gasoline Situation—Methods of Distribution," June 29, 1920; typed chart showing "Basis of Gasoline Delivery as of July 15, 1920."

[17] One contemporary observer espoused this view: "The reason for the failure of gasoline to advance... is a fear of the attitude of Washington" (*National Petroleum News*, March 10, 1920, p. 103).

Some vestiges of the war did remain. Most important was the Lever Food Control Act, which had given the federal government broad powers to control prices and punish "profiteers." Although this law was always on a shaky legal footing, the Justice Department used it to launch an "antiprofiteering" campaign in late 1919 (Eugene Rotwein, 1945, p. 250). But the period's rapid inflation indicates that the prospect of government prosecution did not paralyze many businessmen. Given this, what special circumstances might have made western oilmen feel more threatened or more timid?

In 1919 and 1920, there were several federal investigations affecting the West Coast petroleum industry. The first and most important of these was a Federal Trade Commission inquiry that began in July 1919 in response to complaints made by Puget Sound shipping interests to Senator Miles Poindexter of Washington. While the FTC uncovered some instances of local collusion, it concluded that the "scarcity of crude, resulting from a steady growth of demand" adequately explained the higher prices and earnings between 1916 and 1920.[18] There were no formal complaints, negotiated settlements, or concessions of any form.

The major oil companies treated the ongoing investigation very seriously, but little evidence suggests that it tied their hands. In the spring of 1920, when the fear-of-government hypothesis implies that SOCal would be retreating on all fronts, the company embarked on an aggressive merger bid to take over Midwest Refining, the largest oil company in the Rockies. SOCal's leaders decided to proceed with the clear expectation that this would invite serious legal problems. We also know that during the height of the famine SOCal's president, Kenneth Kingsbury, was not overly fearful, because he wrote to the FTC requesting an early release of the agency's findings "as a matter of justice."[19] SOCal's officials hoped that the report would allay the suspicions of a hostile public.[20]

An examination of other FTC activities considerably strengthens this interpretation. In April 1920, the FTC investigated the nationwide March price increases. While the inquiry was in progress, eastern oil companies advanced gasoline prices again. On June 1, 1920, the FTC concluded that the recent price increases were more likely due to "varying conditions of supply and demand... than to a combination in restraint of trade" and that the oilmen's explanations had "substantial force."[21] It would be hard to imagine a less threatening outcome, and shortly after its release, gas prices advanced yet again in eastern markets.

The U.S. Department of Justice also kept a close eye on the oil industry, again without any formal consequences. The first flurry of activity grew out of the same charges that exercised Poindexter in July 1919. The Department irritated the oil industry by publicizing the shippers' charges in hopes of drawing out complaints that others might harbor. But as the facts became known, the investigation simply slipped from view. This set the pattern for subsequent Justice Department inquiries. Unlike the more thorough FTC studies, these investigations typically became embroiled in the campaigns of the press and special interest groups against the oil industry. With considerable justification, oilmen complained that they were being tried in the press and that the Justice Department pigeonholed favorable findings that would have lessened public hostility.[22] For all of the newspaper indictments, there was only one prosecution. The heavy hand of the law came down on southern California garage owner Homer Smith. He allegedly profiteered by selling gas for $1.00 a gallon (over four times the going rate) to motorists who lined up at his pumps.[23]

[18] White (pp. 552–57); FTC (1921, Part I, pp. 14–15).

[19] SOCal Archives, letter, K. R. Kingsbury to FTC Commissioner N. B. Gaskill, July 20, 1920, Box 120801.

[20] SOCal Archives, letter, Oscar Sutro to K. R. Kingsbury, June 25, 1920, Box 120875.

[21] U.S. House of Representatives (1920, pp. 55–56).

[22] U.S. Department of Justice, "Classified Subjects," and *California Oil World*, July 1; August 12, 1920.

[23] *L.A. Times*, July 23, 1920.

The Justice Department's activities continued through September 1920, but given the massive campaigns to import gas, it is hardly surprising that the ubiquitous claims of hidden storage facilities overflowing with gas proved illusory. Since the oil companies knew that the charges were absurd, they hardly seem a basis for a serious legal threat, although they undoubtedly created a public relations nightmare. It is noteworthy that the major inquiries came after rationing was in effect, and were a consequence rather than a cause of the decision to ration. Although the Justice Department received numerous complaints about oil company policies in other parts of the country, there were no major prosecutions.[24]

A review of other government-industry relations at both the state and federal levels suggests that neither the formal legal environment nor the actions of individual bureaucrats were sufficiently different from those prevailing in other regions at other times to explain why the oil companies in the West rationed in 1920. Investigations were a fact of life in the oil industry. Both before and after 1920, federal and state authorities subjected the industry to periodic scrutiny without inducing radical departures from the market order.

Most federal investigations resulted directly from the attack of groups representing farmers, garage owners, auto dealers, and so on. Special interest group theory suggests that the oil companies might try to buy off a small group of politically dangerous customers at the expense of a larger unorganized population. Even if this were their intent, why would oilmen ration? A large body of economic thought suggests that, without assigning property rights to gas, the queueing and chaos accompanying nonprice rationing would dissipate the rents associated with lower prices, gaining oilmen few friends (Yoram Barzel, 1974; Steven Cheung, 1974). But the allocation program did transfer entitlements to specific users. Station operators often reserved supplies for "established accounts," refusing to serve new buyers, and, more important, oil companies delivered a large percentage of gasoline in bulk to commercial and agricultural users.

The oil companies did not follow a consistent policy if their objective was to buy off one or another group—farmers were among the most favored. But this privileged status came only after weeks of crop-threatening shortages. Despite the subsequent efforts to supply the countryside, many farmers could not carry out normal operations for several months. Furthermore, as a result of explicit policies to increase gasoline supplies, the oil companies cut deeply into the production of engine distillate, the low-cost tractor fuel. SOCal, amidst an uproar of complaints, urged farmers and other users of distillate to convert their engines to burn kerosene. More generally, it is hard to see why the threats of special interest groups were more serious on the West Coast in 1920 than at other times or in other regions. Nor is there any indication that oilmen viewed the allocation scheme in such narrow political terms.

Neat explanations viewing firm behavior as a calculated response to the activities of self-interested government bureaucrats or special interest groups are clearly wanting. This is not to say that the companies were unconcerned with an adverse political fallout. Our reading of the events suggests a different conceptual framework in which these forces operated. Significant increases in the relative price of a good, particularly if it is a "necessity," often provoke howls (James Buchanan and T. N. Tideman, 1974). These usually fall on the deaf ears of the competitive marketplace. But highly visible firms had problems, especially if the price increases could not be justified by unavoidable rises in input costs, and instead led to substantially higher profits. Eastern gasoline marketers, who bought most of their crude from other firms, could attribute their price

[24] In late 1919, U.S. Attorney General Palmer made a number of accusations, threatened prosecutions, and arrested officials of the Transcontinental Oil Company in disputes over fuel oil price increases in the midwestern and southern markets. But the Transcontinental case was dismissed in January 1920, and this activity apparently did little to retard subsequent price increases that year (*National Petroleum News*, December 10, 1919, pp. 30, 75; December 17, 1919, pp. 27–28; December 24, 1919, p. 88).

advances to higher costs, but the integrated western firms could not offer this excuse. Internal documents show that SOCal's leaders worried that their firm's conspicuous position and enormous profits could make it "a target for socialistic or other hostile attack." They were clearly concerned with their public image and tried to maintain the appearance of being "fair."[25]

Sensitivities to fairness and the willingness to accept the legitimacy of the market undoubtedly vary over time and are influenced by recent experiences, education, and ideology. The perceived causes of the shortage also contributed to the "logic" of rationing. To the extent that the gas famine was seen as the result of the drought (an act of nature) and the illegal railroad walkout (an act against society), it would be improper for a "bastion of capitalism" to exploit the situation. The recent wartime experiences also strengthened the sentiment for nonprice rationing as a patriotic way to deal with market adjustments.

Paradoxically, the more severe shortfall in the West created a crisis atmosphere, making it more difficult to rely on the market. By the early spring a 2-to-3-cent price increase almost certainly would not have cleared the market as it did in the East. Given existing interregional price differences, transportation difficulties, and reasonable guesses about demand elasticities, price increases of this magnitude would have done little to relieve the situation in the short run. For SOCal to increase its prices a few cents without visibly solving the problem would have been seen as callously profiting from public misfortune instead of sharing the burden. To clear the market might have required up to a doubling of prices creating a *cause célèbre* that any democratic government would find difficult to ignore. In the emotionally charged climate, movements for more extreme sanctions could have gained momentum especially if established avenues for discontent, such as federal inquiries, had been closed off. On several occasions oilmen actually sought and received the help of government officials to offset inflammatory rumors and newspaper accounts. Even if political institutions had been more favorable to oil companies, oilmen might have feared a fundamental change in the political environment brought on by a groundswell of hostility.

[25] SOCal Archives, "Confidential Memorandum" drafted by Oscar Sutro, May 14, 1920, Box 120775, and *Standard Oil Bulletin*, July 1920.

VII. Conclusion

Much of the reaction to the gasoline crises of the 1970's sprang from the belief that such shortages had never before been experienced in peacetime. Beyond correcting this widely held view, a recounting of the 1920 shortage raises important questions about the conclusions drawn from the more recent past. Most analysis measures the actual events under a regulated regime against the counterfactual of the competitive marketplace. But in 1920 a combination of economic, social, and political forces dulled the market's response. Private firms, exhibiting an extraordinary amount of market power, voluntarily assumed a role normally attributed to governments. Might similar forces have temporarily blunted market responses in the 1970's in the absence of government controls?

Our inquiry into why the leading oil companies decided to ration fails to produce a "smoking gun," but our comparative analysis does suggest that there are inconsistencies in several versions of the government hypothesis. We have focused on the decision not to raise prices—after all, the standard paradigm views most shortages as market failures resulting from price controls. This led us to questions such as, were oil executives afraid of future hostile government actions, or, more basically, public hostility to price increases? But there is a different perspective, one which we expect may have dominated oil company boardrooms. The rationing program was framed to allocate resources that were becoming scarce to specific sets of users. Although there was squabbling as to who qualified at the margin for favored treatment, the makeup of the intramarginal membership enjoyed widespread consent. For example, the granting of

preferential treatment to farmers whose crops needed harvesting was not open to question. The executives of the larger oil firms probably shared this view, but in any case they understood it. Whether for reasons of social responsibility or out of a fear of hostile reactions, they designed administrative programs to allocate supplies rather than trusting the marketplace. Again there was a consensus. We did not find any commentary in the printed medium or in firm records even suggesting that prices should be used to allocate gas during the shortage. It was only in July, after the highly publicized relief trains arrived from Oklahoma, that some observers asked why higher prices had not been used to attract fuel imports earlier.

According to SOCal's past chairman, R. G. Follis, oilmen also viewed the purely economic consequences of a large, market-clearing price increase as detrimental to their long-run interest.[26] SOCal's leaders saw their company's prosperity as integrally tied to the economic development of the West, and they accepted considerable responsibility for promoting that development. Given this attitude, Follis thought that it would have been unwise to shock the economy with enormous fluctuations in oil prices. To encourage western economic development and the rapid conversion to petroleum fuels, industry leaders thought it essential to assure agriculture and business a guaranteed supply of energy. The preference system had this explicit aim.

The 1920 gas famine would be far less troubling if it were a single aberration, but it is not. Our cursory examination of the U.S. energy market has turned up two other cases of widespread private rationing—during the coal famine in the winter following the 1902 miners' strike and during the "hold-that-price" gasoline shortage of 1947 and 1948.[27]

These events deserve further inquiry. The analysis of instances where markets fail to work as anticipated should provide us with a better perspective of the robustness of our standard micro models, and help us understand the delicate interaction between economic, political, and ideological variables needed for accurate prediction.

[26] Interview with R. G. Follis, San Francisco, October 27, 1981.

[27] When the market price of coal climbed to $8 or $9 a ton in 1902, major distributors held firm at about $5. Newspapers reported widespread rationing during the bitterly cold winter. A few of the highlights noted in the *New York Times* are: "Men and women fought with one another to get a chance to buy a beggarly pail of coal.... Hundreds had to be turned away.... No one person was permitted to purchase more than two pails" (December 10, 1902); "After taking one's place in line it required about two hours of slow trudging and waiting before the coveted thirty pounds of fuel could be purchased" (December 11, 1902); "Bankers and Preachers Seize A Coal Train" (January 11, 1903). For a discussion of 1948, see Paul Giddens (1955, pp. 680–85); Harry Hansen and Powell Niland (1952, pp. 114–32); Henrietta Larson et al. (1971, pp. 664–85).

REFERENCES

Barzel, Yoram, "A Theory of Rationing by Waiting," *Journal of Law and Economics*, April 1974, *17*, 73–96.

Buchanan, James M. and Tideman, T. N., "Gasoline Rationing and Market Pricing: Public Choice in Political Economy," *Atlantic Economic Journal*, November 1974, *2*, 15–26.

Cheung, Steven N. S., "A Theory of Price Control," *Journal of Law and Economics*, April 1974, *17*, 53–72.

Friedman, Milton, "Blaming the Obstetrician," *Newsweek*, June 4, 1979.

Giddens, Paul H., *Standard Oil Company (Indiana): Oil Pioneer of the Middle West*, New York: Appleton-Century-Crofts, 1955.

Hansen, Harry L. and Niland, Powell, "Esso Standard: A Case Study in Pricing," *Harvard Business Review*, June 1952, *30*, 114–32.

Hopkins G. R., "Petroleum Refinery Statistics 1916–1925," in U.S. Bureau of Mines, Bulletin 280, Washington, 1927.

Larson, Henrietta M., Knowlton, Evelyn H., and People, Charles S., *History of Standard Oil Company (New Jersey): New Horizons, 1927–1950*, New York: Harper & Row,

1971.

Nash, Gerald D., *United States Oil Policy, 1890–1964,* Pittsburgh: University of Pittsburgh Press, 1968.

Owen, Edgar W., *Trek of the Oil Finders: A History of Exploration for Petroleum,* Tulsa: American Association of Petroleum Geologists, 1975.

Rotwein, Eugene, "Post-World War I Price Movements and Price Theory," *Journal of Political Economy,* December 1945, *53,* 234–57.

Welty, Earl M. and Taylor, Frank J., *Black Bonanza,* New York: McGraw-Hill, 1958.

White, Gerald T., *Formative Years in the Far West: A History of Standard Oil Company of California and Predecessors Through 1919,* New York: Appleton-Century-Crofts, 1962.

Federal Trade Commission, *Pacific Coast Petroleum Industry;* Part I: *Production, Ownership, and Profits,* Washington: USGPO, 1921.

______, *Pacific Coast Petroleum Industry;* Part II: *Prices and Competitive Conditions,* Washington: USGPO, 1922.

______, Record Group 122, Boxes 3317–45, Washington National Records Center, Suitland, MD.

Standard Oil Company of California (SOCal) Archives, "Extracts from Salesmen's and Special Agent's Daily Reports, July 1920," Box 120775; Sales Department, "Annual Reports, 1891–1922," typed ms., Box 120762; sundry confidential letters, telegrams. and memos, Boxes 120786, 120801, 120803, 120875, 120876, Dublin, CA.

Union Oil Company of California Archives, "Report of the Manufacturing Department, 1920," Appendix B, Refinery Operations, typed ms. and "Annual Reports," Santa Fe Springs, CA.

U.S. Department of Justice, Attorney General's Office, Record Group 60, "Classified Subjects," Box 329, 60-57-10, Sec. 11, National Archives, Washington, D.C.

U.S. Geological Survey, *Mineral Resources of the United States 1921;* Part II: *Non-Metals,* Washington: USGPO, 1924.

______, *Mineral Resources of the United States 1922;* Part II: *Non-Metals,* Washington: USGPO, 1925.

U.S. House of Representatives, 66th Congress, 2nd Session, DOC. 801, "Advances in the Price of Petroleum Products," Washington: USGPO, 1920.

Monetary Policy Games and the Role of Private Information

By MATTHEW B. CANZONERI*

Consider the following scenario: monetary aggregates surge ahead of expected or targeted rates; the Fed claims that it is accommodating a perceived increase in money demand in order to stabilize the price level; the private sector (or the administration, or Congress) counters that the Fed is running an inflationary policy to expand employment; a period of credibility building ensues, which may focus upon the personalities of policymakers, targeting procedures, or even proposals for legislative reform. A key element in this scenario is that the private sector cannot verify the Fed's claim; the Fed's forecast of money demand is private information.

A plausible explanation for such a scenario is that the Fed is involved in a noncooperative game with certain agents in the private sector, or with some other branch of government. Why else, for example, would the Fed seek to constrain its own behavior, with say a targeting procedure, unless it thought that this would bring a resolution to a credibility problem resulting from a noncooperative game.[1] And why does the Fed have a "credibility" problem, as opposed to an "information" problem; that is, why are its announcements not accepted at face value?

A game suggested by Finn Kydland and Edward Prescott (1977) has received considerable attention.[2] In this game, wage setters have to specify the nominal wage in a labor contract before the Fed sets the money supply. The Fed wants a higher level of employment than the wage setters, but it is also inflation conscious. The wage setters know that the Fed will be tempted to inflate away some of their real wage, to achieve a higher level of employment, so they purposely set the wage high. The Fed weights its employment and inflation objectives, and only finds it optimal to inflate the real wage down to the level wanted by the wage setters. The noncooperative solution has an inefficient inflation bias, simply because the Fed has no credible way of precommitting itself to a noninflationary policy. Adding a stabilization role for monetary policy, and private information, the Kydland-Prescott game can be used to model the scenario outlined above.

There are a variety of ways in which the noncooperative solution described above might be improved upon. Congress could legislate institutional reforms. It could legislate a monetary policy rule directly, or it could constrain the behavior of the players by changing the rules of the game; wage-price guidelines and legislated targeting procedures are examples of the later. Earl Thompson (1981) and Kenneth Rogoff suggest changing the incentive structure for Fed policymakers, or simply choosing a policymaker with "perverse" preferences. And Robert Barro and David Gordon (1983a) suggest that the players themselves may have already found a resolution to the problem in a reputation-building mechanism.

An efficient resolution of the credibility problem must leave the Fed with the latitude to perform its stabilization role. Rogoff and

*Georgetown University, Washington, D.C. 20057. I thank R. Cooper, J. Friedman, E. Green, D. Henderson, B. McCallum, K. Rogoff, J. Taylor, and an anonymous referee for helpful comments; however, the views presented here are my own. They are not necessarily shared by the Federal Reserve Board or any other member of its staff.

[1] Targeting money or interest rates is not likely to be efficient in terms of meeting final targets; see, for example, my 1977 article. This observation is the principal theme of Kenneth Rogoff (forthcoming).

[2] See Robert Barro and David Gordon (1983a,b), John Taylor (1983), S. Green (1983), Rogoff, David Baccus and John Driffill (1985), and more generally, W. Lee (1981). Edmund Phelps (1967) had many of the elements of this game.

Barro and Gordon (1983a) discuss the tradeoff between flexibility, that is required for stabilization, and constraint, that is required for a resolution of the credibility problem, but neither of their papers incorporates private information. An efficient resolution is much more difficult to achieve if the Fed's forecast of money demand is indeed private information. No stable resolution of the credibility problem can rely on the Fed's own announcement of its forecast if the Fed has an incentive to misrepresent; the resolution must be, in Leonid Hurwicz's (1972) terminology, incentive compatible.[3]

The present paper proceeds as follows. Section I outlines the Kydland-Prescott game in some detail and describes the modifications necessary to give the Fed a stabilization role. The importance of private information is highlighted. Without private information, John Taylor's observations seem justified; resolutions of the precommitment problem would be too easy to come by for this game to be taken seriously as a real world problem.

Section II reviews the work of Rogoff and Barro-Gordon (1983a) in the light of private information. Rogoff's theory of the perverse policymaker may be less attractive in the presence of private information since the policymaker's actions do not reveal his true preferences. However, the incentive compatibility problem that results may afford a compelling explanation of the sometimes dramatic comings and goings of individual policymakers. The rest of Section II (and Section III) focuses exclusively on repeated-game solution concepts. In a repeated game, honest behavior in the present period can be make to yield rewards or expanded opportunities in the next. Thus, there is an added strategic dimension; a player can invest in his reputation. Using James Friedman's (1971) notion of reversionary periods, Barro-Gordon (1983a) develop a model that resolves the credibility problem in the absence of private information. Their model falls apart if the Fed's forecast of money demand is private, because wage setters cannot tell when the Fed has been honest. However, as Edward Green and Robert Porter (1984) have noted in a different context, this may be a blessing in disguise. The Friedman solution, like Rogoff's, is too stable to explain the episodic periods of breakdown that are actually observed. A reformulation of the model accommodates private information and actually predicts periodic bouts of inflation (though it would be inappropriate to think of these episodes as credibility breakdowns). It also suggests that, despite these periodic difficulties, the Fed and the wage setters may have already found a rather satisfactory resolution to the credibility problem.

Section III discusses the legislative approach in a more sophisticated way than was possible in Section I; the proper legislation can also make the Fed's opportunities in future periods depend upon honest behavior in the present. There are monetary policy rules that produce a better game outcome (generally at some expense to the stabilization effort) and are incentive compatible, even if the Fed's forecast of the money demand is private information. These rules are, however, less intuitive than the typical stabilization rule; they are more easily motivated as the Fed's optimal response to a congressionally mandated targeting procedure.[4] Various targeting procedures are considered, some leaving the Fed with more latitude than others. Which is best depends on the Fed's ability to predict money demand disturbances and on the importance society attaches to the benefits of stabilization relative to the costs of the inflation bias resulting from the noncooperative behavior of the players. Section IV summarizes the basic results of the paper and suggests where further research is needed.

[3] This paper borrows heavily from Hurwicz, as does much of the recent contracting literature; see Robert Townsend (1982), Oliver Hart (1983), and the *Quarterly Journal of Economics* (suppl. 1983).

[4] They could also be motivated along the lines of Peel's Bank Act of 1844; see Axel Leijonhufvud's 1982 discussion of the Peel-Friedman system.

I. The Game and the Importance of Private Information

Consider a wage contracting model like those of JoAnna Gray (1976) and Stanley Fischer (1977a). The basic assumption of this model is that the nominal wage must be set in a labor contract prior to the setting of the money supply, and thus the realization of the price level. The wage is set at a level that is expected to clear the market. Later, during the contract period, firms learn the price from the output market, and the contract provides that they maximize profits; they hire along their marginal productivity curve.[5] The output supply function can be expressed as

$$(1) \qquad y_t = \bar{y} + \theta(p_t - w_t),$$

where y_t, p_t, and w_t are the logs of output, the price level, and the contract wage; $\bar{y}$ is the equilibrium rate of output, corresponding to the "natural" rate of employment.

Wage setters want to achieve the equilibrium output y. Their utility can be represented by[6]

$$(2) \qquad UW_t = -(p_t - w_t)^2 = -(1/\theta^2)(y_t - \bar{y})^2.$$

Of course, p_t is not known when the contract is made; the expected utility-maximizing strategy for wage setters is thus

$$(3) \qquad w_t = p_t^e,$$

where p_t^e is the wage setters' prediction of the price level that will obtain in the contract period. The wage setters' play is, essentially, their prediction.

Substituting (3) into (1) results in a familiar prediction error model of supply. Note also that (1) and (2) can be expressed in terms of inflation and expected inflation:

$$(4) \qquad y_t = \bar{y} + \theta(p_t - p_t^e) = \bar{y} + \theta(\pi_t - \pi_t^e)$$

$$(5) \qquad UW_t = -(\pi_t - \pi_t^e)^2,$$

where $\pi_t = p_t - p_{t-1}$ is the actual rate of inflation, and $\pi_t^e = p_t^e - p_{t-1}$ is expected inflation.

Equation (2) (or (5)) gives the wage setters' utility. The Fed maximizes social utility, which is given by

$$(6) \qquad US_t = -(y_t - k\bar{y})^2 - s(\pi_t - \pi^*)^2,$$

where $k > 1$, and π^* is the optimal rate of inflation. The rate of output sought by wage setters, $\bar{y}$, is too small from a social point of view.[7] This assumption can be motivated in several ways. My 1980 paper assumes that the labor market is dominated by large unions. The labor supply curve includes only union members, and wage setters' behavior systematically excludes other workers. By contrast, the social utility function includes all workers. Barro and Gordon (1983a, b) have a different view. They assert that tax policy and unemployment compensation distort the labor supply and lead to a socially inefficient equilibrium. (Under this interpretation, equation (2) is a guide to wage setters' behavior, but not a measure of their welfare.) Either view can be used to rationalize the conflict between (2) and (6) over optimal levels of the real wage, employment, and output.[8] The second term in the social utility function penalizes deviations from the optimal rate of inflation. In a less

[5]This assumption has been criticized; see the exchange between Barro and Fischer (1977b) and also D. Waldo (1981).

[6]Joshua Aizenman and Jacob Frenkel (1983) provide a rationale for this utility function. The analysis is simplified by abstracting from productivity disturbances; see Gray and Rogoff.

[7]It is not necessary to specify conflicting utility functions in order to generate the time inconsistency on which this game rides; see Fischer (1980) and my forthcoming monograph with Dale Henderson for expositions of the time consistency problem that are mathematically unencumbered.

[8]While the two views are equivalent for present purposes, they do have different implications for the first-best solution to the policy problem posed here. My 1980 view suggests labor legislation, while Barro and Gordon's (1983a, b) points to fiscal policy reform.

than fully indexed economy, inflation redistributes income from one group to another and distorts incentives; in addition, there is seigniorage. Many factors go into the calculation of π^*; here, π^* is rather cavalierly taken to be a regime independent constant.

Some will prefer not to identify the Fed's utility function (6) with social utility. In this view, (6) merely reflects the biases of Fed policymakers. It is sometimes asserted, for example, that macro policymakers have tended to underestimate how high the "natural rate" of unemployment has actually been, and that while the "evils" of inflation are poorly documented, there is clearly a political mandate to sacrifice some employment and output to keep inflation in check. If this view of (6) and its conflict with (2) is adopted, the normative content of what follows must be appropriately reinterpreted.

To explain the Fed's interaction with the wage setters, the rest of the economic environment must be modeled. This is done in the most rudimentary way; that is, control theoretic aspects are trivialized so as to focus upon game theoretic aspects. The model is closed with a simple quantity equation,

$$m_t - p_t = \bar{y}, \tag{7}$$

where m_t is the log of the money supply. First differencing, (7) becomes

$$g_t - \pi_t = 0, \tag{8}$$

where $g_t = m_t - m_{t-1}$ is the growth rate of the money supply; g_t is taken to be the Fed's instrument.

The basic structure of the precommitment problem should now be apparent. The wage setters must act first; this is an assumption of the contracting model. They must specify w_t in labor contracts, not knowing what g_t will be, but fully aware of the Fed's motivations as embodied in the social utility function. The wage setters know that the Fed will be tempted to inflate away their real wage and achieve its higher output target, $k\bar{y}$. The wage setters can make this strategy expensive for the Fed, in terms of its inflation goal, by setting w_t high in the first place; in fact, if they set w_t high enough, the Fed will find it optimal to give them their preferred real wage and output, $\bar{y}$. This is the problem analyzed in my 1980 paper, and by Barro-Gordon (1983a), and by Kydland-Prescott before them. As they all point out, the noncooperative solution to this game is inefficient. If the Fed could somehow be credibly committed to setting g_t equal to π^*, then wage setters would not feel compelled to set a high rate of wage inflation. Wage setters would still get their desired rate of output, $\bar{y}$, but society would be spared the unnecessary inflation.

The message so far is that the inefficiencies brought about by the conflict over output goals and the timing of players' actions could be eliminated by tying the hands of the Fed. This would be the final message if there were no benefit to society from the Fed's retaining a degree of flexibility. There is just such a benefit to society if the Fed can play a stabilization role.

Suppose a stochastic disturbance is added to money demand in (7). If this disturbance follows a random walk, then (8) becomes

$$g_t - \pi_t = \delta_t, \tag{9}$$

where δ_t is the white noise innovation. Suppose further that wage setters do not see δ_t at the time they have to specify w_t, but that the Fed does have δ_t (or some forecast of it) when it sets g_t. More specifically, let δ_t be decomposed into e_t and ε_t,

$$\delta_t = e_t + \varepsilon_t, \tag{10}$$

where e_t is the Fed's forecast of δ_t, and ε_t is its forecast error. If the Fed is left with sufficient flexibility, it can accommodate e_t before it passes on to prices and output; this benefits both wage setters and society as a whole. The problem now is one of trading off the flexibility needed for stabilization with the constraint required for eliminating the inflation bias; this problem is touched upon in Barro and Gordon (1983a), and it is the main subject of Rogoff.

Two quite different information structures will be considered below, and it is convenient to be more explicit about them at this point. With "symmetric information," both

wage setters and the Fed are assumed to observe e_t at the same time, though it is too late for the wage setters to incorporate this new information into the contract wage, w_t. With "private information," either the Fed has superior information that cannot be revealed or the policymaking process is such that the private sector cannot reconstruct the Fed's forecast of money demand. At the end of the contract period, wage setters can calculate δ_t from (9), for they will then observe both g_t and π_t. However, they cannot decompose δ_t into e_t and ε_t. The Fed's forecast, e_t, is private information.

Before going on, it is also convenient to express the players' utility functions in terms of their actions, and to give a more formal description of the game. Using (9), the wage setters' utility function (5) can be expressed:

$$UW_t = -(g_t - g_t^e - \delta_t)^2, \tag{11}$$

where g_t^e is the wage setters' prediction of the Fed's action at the time they must set w_t. In view of (3) and the fact that $g_t^e = \pi_t^e$, it should be clear that g_t^e can be taken to be the wage setters' action. Using (9) and (1), the social utility function (6) can be expressed as

$$UF_t = -(g_t - g_t^e - \delta_t - y^*)^2 - f(g_t - \delta_t - \pi^*)^2, \tag{12}$$

where $UF_t = US_t/\theta^2$, $f = s/\theta^2$ and $y^* = (k-1)\bar{y}/\theta$; θy^* is the difference between the players' output goals. The game proceeds as follows: 1) wage setters form g_t^e on the basis of their knowledge of the Fed's utility function and their expectation that δ_t (and e_t and ε_t) is zero. 2) The Fed sets g_t to maximize its expected value of UF_t in (12), knowing g_t^e and its own forecast of δ_t. 3) *Ex post*, wage setters observe δ_t, and if information is symmetric, they can decompose it into the Fed's forecast, e_t, and the forecast error, ε_t. If the Fed's forecast is private information, they cannot make this decomposition.

The noncooperative solution can now be found by working backwards through the sequence of decisions that the players have to make. First, maximize EUF_t taking g_t^e as a fixed parameter; the first-order condition gives the Fed's decision rule as a function of the wage setters' action, g_t^e. Then take the expectation of the Fed's decision rule (conditional on information available to the wage setters when they set w_t) to find the wage setters' utility-maximizing action, $g_t^{e,nc}$. The resulting *Noncooperative Solution* is

$$\begin{aligned} (13)\quad & g_t^{nc} = \pi^* + e_t + y^*/f, \\ & g_t^{e,nc} = \pi^* + y^*/f, \\ & y_t = \bar{y} - \theta\varepsilon_t, \qquad \pi_t = \pi^* + y^*/f - \varepsilon_t, \\ & EUF_t^{nc} = -(1+f)\sigma_\varepsilon^2 - [1+(1/f)]\,y^{*2}, \\ & EUW_t^{nc} = -\sigma_\varepsilon^2. \end{aligned}$$

This is a Nash solution; neither player can expect to do better, given the strategy of the other. And the solution holds under either information structure; that is, one obtains the same noncooperative solution whether or not the Fed's forecast of money demand is private information.

Wage setters get exactly what they want in this solution. They set wage inflation high enough that the Fed is not tempted to try to achieve its higher output goal, and the Fed fully accommodates predictable money demand disturbances, stabilizing output to the fullest extent possible about the wage setters' preferred rate. There is, of course, an inflation bias equal to y^*/f. Society is left (on average) with a level of employment that is too low and a rate of inflation that is too high; as Barro-Gordon (1983b) and Rogoff point out, this is a model of stagflation.

Now, it may be thought that the noncooperative solution just described is a rather silly, destructive situation for the Fed and the wage setters to allow themselves to fall into. The players have conflicting output goals, but the inflation bias benefits no one. An *Ideal Solution* would eliminate the inflation bias without changing the (average) rate

of output;

$$(14)\qquad g_t^c = \pi^* + e_t, \quad g_t^{e,c} = \pi^*,$$
$$y_t = \bar{y} - \theta\varepsilon_t, \quad \pi_t = \pi^* - \varepsilon_t,$$
$$EUF_t^c = -(1+f)\sigma_\varepsilon^2 - y^{*2} > EUF_t^{nc},$$
$$EUW_t^c = EUW_t^{nc}.$$

Here, the Fed fully accommodates the predictable part of the money demand, and it does not try to inflate away θy^*. The solution is efficient in the sense that no other set of strategies, g_t and g_t^e, can make one player better off without making the other worse off; thus the Ideal Solution can be thought of as a cooperative solution. (That is why the superscript c is used to identify it.)

As Kydland and Prescott, Barro-Gordon (1983a), and my 1980 paper have all pointed out, the problem with this solution is that the Fed's policy is time inconsistent. Once the wage setters have committed themselves to $g_t^{e,c}$, g_t^c will no longer be the Fed's utility-maximizing action. Letting $g_t^e = \pi^*$ in (12), and maximizing EUF_t with respect to g_t, the best *Cheating Solution* for the Fed is

$$(15)\qquad g_t^{ch} = \pi^* + e_t + y^*/(1+f),$$
$$g_t^{e,ch} = \pi^*,$$
$$EUF_t^{ch} = EUF_t^c + y^{*2}/(1+f),$$
$$EUW_t^{ch} = EUW_t^c - y^{*2}/(1+f).$$

Wage setters cannot set w_t on the expectation that inflation will equal π^* and then count on the Fed to follow through with g_t^c. The Fed will be tempted to inflate a little more than π^*, lowering the real wage and inflating away some of θy^*. Unless the Fed can be credibly precommitted to the ideal policy, wage setters will almost surely play noncooperative.

Can the Fed be credibly committed to the ideal policy at the time that wage setters have to set w_t in their contract? Views seem to differ as to whether this is a real world problem. Taylor, commenting on Barro-Gordon (1983a,b), points out that society has found sensible ways out of other obvious time inconsistency problems; he cites the example of patent law. Taylor seems skeptical that the Kydland-Prescott game really captures the credibility problem the Fed is generally thought to face.

My own view is that Taylor would probably be right were it not for private information. Absent private information, resolutions of the precommitment problem would be relatively easy to find. Suppose information is symmetric, and wage setters observe e_t at the same time the Fed does. The Fed might be able to establish credibility in the ideal policy by simply running it for a number of periods. If wage setters actually observe e_t each period, they can verify that the Fed is implementing the policy g_t^c, and they might well be willing to go along.[9] Alternatively, Congress could legislate the ideal policy rule and police the Fed's adherence to it.[10]

If the Fed's forecast of money demand is private information, a resolution of the precommitment problem is much more difficult to come by. If e_t is not observed by the wage setters or Congress, then direct verification of the Fed's adherence to the ideal policy rule is not possible. The Fed is often urged to announce a forecast, e_t^a, at the time it sets g_t. However, this would be of little use in the present context, for in the language of Hurwicz, the mechanism would not be incentive compatible; the Fed would have an incentive to misrepresent its information.[11] By letting its announcement e_t^a be equal to the true forecast e_t, plus $y^*/(1+f)$, the Fed

[9]Some experimental studies suggest that players may be willing to go along in this way; see Alvin Roth and Francoise Schoumaker (1983).

[10]In a more realistic setting, the ideal policy rule could well be more complicated, but this should not in and of itself create a credibility problem. Credibility problems arise only when scope has been left for cheating.

[11]Wage setters can tell if the Fed is telling the truth "on average," for then the sum $\Sigma(\delta_t - e_t^a) = \Sigma\varepsilon_t$ should tend to zero. If the relationship is an enduring one, and if utility is not discounted over time, then this approach might work. See Roy Radner (1981).

can make the cheating policy (15) look like the ideal policy; that is, $g_t^{ch} = \pi^* + e_t^a$. Straightforward reputational or legislative approaches to the problem seem more likely to result in the scenario outlined in the introduction. The Fed sets a high g_t, claiming to be accommodating an unexpectedly high money demand; the private sector charges that it is instead running an inflationary policy to expand output, and a credibility breakdown ensues.

The results of this section may be summarized as follows. The Kydland-Prescott game of precommitment has a simple resolution if there is no benefit to the Fed's retaining stabilization powers; the Fed's hands should be tied with a k-percent rule. If the Fed can play a useful stabilization role, then an efficient resolution of the precommitment problem requires that the Fed should retain some flexibility. However, straightforward resolutions still exist if information is symmetric; the Fed might be able to build a reputation for itself by simply running the ideal policy for a few periods, or Congress could legislate the ideal policy rule and police the Fed's adherence to it. If the Fed's forecast of money demand is private information, then the Fed's adherence to the ideal policy rule cannot be verified directly, and efficient resolutions of the precommitment problem are much more difficult to find.

II. Adding Private Information to Rogoff and Barro-Gordon

Rogoff and Barro-Gordon (1983a) discuss more sophisticated approaches to the precommitment problem. However, neither allows for private information. If the Fed's forecast of money demand is private information, both approaches yield more interesting interpretations of the credibility problem, but Rogoff's does not offer a resolution and Barro-Gordon's does not explain the scenario outlined in the introduction.

A. *The Perverse Policymaker*

It is sometimes asserted that monetary policymakers have a conservative or anti-inflation bias, or that they have an unnatural proclivity towards monetary targets. Rogoff suggests that this might be one of society's ways of trying to improve upon the Noncooperative Solution. Society may have found that choosing a policymaker with preferences other than those expressed in (12) leads to a noncooperative solution that is closer to the Ideal Solution.

Suppose the Fed is run by a perverse policymaker with the utility function

$$(16)\quad UP_t = -(g_t - g_t^e - \delta_t - y^*)^2 - \phi f(g_t - \delta_t - \pi^*)^2.$$

UP_t puts more weight on inflation than UF_t if $\phi > 1$. The *Noncooperative Solution with a Perverse Policymaker* is

$$(17)\quad g_t^p = \pi^* + e_t + y^*/\phi f,$$

$$g_t^{e,p} = \pi^* + y^*/\phi f,$$

$$EUP_t^p = -(1+\phi f)\sigma_\varepsilon^2 - [1 + (1/\phi f)]\, y^{*2},$$

$$EUF_t^p = EUF_t^c - y^{*2}/\phi^2.$$

The more perverse is the policymaker (i.e., the bigger is ϕ), the closer is solution (17) to the optimal solution (14). The policymaker's sensitivity to inflation allows the wage setters to get their way with less inflation.

Here again, adding private information both complicates and enriches the model. If e_t were directly observable at the end of the period, then wage setters could verify that the policy g_t^p had indeed been implemented, and that UP_t truly represented the policymaker's preferences, rather than UF_t. The policymaker's actions would reveal his preferences.

If e_t is private information, this is not possible. The policymaker could announce a forecast e_t^a when he set g_t, but once again the mechanism would not be incentive compatible.[12] To see this, suppose the policymaker represents his preferences as UP_t (with $\phi > 1$), but actually his preferences are UF_t.

[12] Footnote 11 applies here as well.

Then, substituting $g_t^e = \pi^* + y^*/\phi f$ and $g_t = \pi^* + e_t^a + y^*/\phi f$ into (12) and maximizing EUF_t with respect to e_t^a, the policymaker's optimal misrepresentation $e_t^a - e_t$ can be seen to be $[(\phi - 1)/\phi(1+f)]y^*$. It can be shown that a marginal increase in ϕ over unity will raise EUF_t above EUF_t^{nc}, and lower EUW_t below EW_t^{nc}.

Thus a policymaker with true social preferences has an incentive to look conservative, and if the Fed's forecast of money demand is private information, wage setters have no way of telling for sure. The policymaker's actions and announcements will not reveal his actual preferences. The implications of this observation would appear to be twofold. The model may now provide a more plausible interpretation of history; without private information (and some uncertainty about policymakers' true preferences), Rogoff's approach to the credibility problem is too stable to explain the comings and goings of individual policymakers, with all of the attendant hoopla, posturing, and changes in public perception. However, the very same instability that may help explain history also detracts from the appeal of this approach as a final resolution of the credibility problem.

Rogoff does not consider private information, but he does offer an alternative interpretation of his formal analysis that may be helpful in addressing the issue. He suggests that the added weight on inflation in (16) may be viewed as a commitment to targeting; in this case, the policymaker commits himself to targeting the inflation rate. Rogoff also suggests that the policymaker's commitment should be enforced through a system of rewards and punishments; Congress, for example, might be able to achieve an institutional reform by legislating appropriate bureaucratic incentives. If Fed forecasts are indeed private information, then incentive compatibility is one consideration that will make the design of such a system of rewards and punishments more difficult, but the general approach seems worth pursuing. The targeting discussion in Section III may be viewed in this light.

With this alternative interpretation, Rogoff also has a plausible explanation for the existence of other forms of targeting. Suppose the policymaker's perversity is a penchant for targeting g_t on π^*, so that

$$(18)\quad UP_t = -(g_t - g_t^e - \delta_t - y^*)^2 - f(g_t - \delta_t - \pi^*)^2 - \tau(g_t - \pi^*)^2.$$

In this case, the EUP_t maximizing policy is

$$(19)\quad g_t^p = \pi^* + [(1+f)/(1+f+\tau)]e_t + y^*/(f+\tau),$$

and τ can be chosen to maximize EUF_t.

The size of τ determines the tightness of the targeting procedure. Note that if $\tau > 0$, the policymaker will not try to fully accommodate money demand disturbances. The choice of τ must trade off the benefits of reducing the inflation bias with the costs of constraining the stabilization effort. The choice of ϕ in the last example would involve a similar calculation if productivity disturbances were added to the model. Rogoff considers the relative merits of money, interest rate, inflation rate, and nominal income targeting, all in a much more interesting control setting than was practical here.

B. *A Reputation-Building Approach*

The credibility problem in the Kydland-Prescott game is one of precommitment; the Fed would like to be able to promise the wage setters that it will run the ideal policy, and then have no reason for reneging on its promise later. Barro and Gordon show that Friedman's solution concept for repeated games offers a resolution to this problem.

Suppose the Fed and the wage setters have arrived at the following setup. Wage setters believe the Fed's promise to run the ideal policy in the current period if the Fed did not renege on its promise last period. (Assume for the moment that there is no private information.) If, however, the Fed reneged last period, then wage setters revert to their noncooperative wage settings this period; in subsequent periods, they again believe the Fed's promise, until shown otherwise. Barro and Gordon show that the wage setters are

not irrational in going along with this setup, because the Fed will never have an incentive to renege.[13]

To see this, note that the Fed's temptation to renege in period t is measured by

$$(20) \quad EUF_t^{ch} - EUF_t^c = y^{*2}/(1+f),$$

while its reward in period $t+1$ for not reneging in period t is

$$(21) \quad EUF_{t+1}^c - EUF_{t+1}^{nc} = y^{*2}/f.$$

Since the reward for honesty outweighs the temptation to cheat, it pays the Fed to invest in its reputation; it has no incentive to renege.[14]

There are two weaknesses in this resolution of the credibility problem that can be dealt with in ways suggested by Green and Porter. (There are some others that are not so easily remedied; they will be discussed at the end of the section.) The first weakness is that the Barro-Gordon solution is, like Rogoff's, too stable. If taken literally, it implies that one should never expect to see reversions to inflationary periods. The second weakness is that if the Fed's forecast of money demand is private information, then the wage setters can never know for sure when the Fed has reneged. The wage setters do not observe the Fed's forecast at the end of the period, and as was shown in Section I, it would do no good to have the Fed announce a forecast, for it would not be believed.

The model can be reformulated to deal with both weaknesses as follows. The wage setters do see the money demand disturbance δ_t at the end of the period; they just can't decompose it into the Fed's forecast e_t, and the residual ε_t. Thus, while it is not feasible to have the wage setters revert to their inflationary wage settings when g_t is greater than $g_t^c = \pi^* + e_t$, it is feasible to have a reversionary period triggered if g_t is greater than $\pi^* + \delta_t + \bar{\varepsilon}$, where $\bar{\varepsilon}$ is some appropriately chosen constant. The probability of a reversion in period $t+1$ is[15]

$$(22) \quad P(g_t - g_t^c - \bar{\varepsilon}) = Pr\left[\varepsilon_t < g_t - g_t^c - \bar{\varepsilon}\right],$$

where $P(\cdot)$ is the cumulative distribution function of the Fed's forecast error, ε_t. If the Fed reneges and sets g_t greater than g_t^c, it increases the probability of a reversion in period $t+1$. Similarly, decreasing the constant $\bar{\varepsilon}$ increases the probability of a reversion. Wage setters are not irrational in going along with this setup if $\bar{\varepsilon}$ is small enough that the Fed has no incentive to renege.[16]

The problem now is to find an $\bar{\varepsilon}$ such that if the Fed considers raising g_t marginally above g_t^c, it will conclude that the expected gain in period t is offset by the expected loss in period $t+1$. Suppose g_{t-1} was less than $\pi^* + \delta_{t-1} + \bar{\varepsilon}$ (so that the Fed is not currently in a reversionary period), and let $U(g_t, \bar{\varepsilon})$ be the Fed's expected utility over the next two periods;

$$(23) \quad U(g_t, \bar{\varepsilon}) = EUF_t + PEUF_{t+1}^{nc} + (1-P)EUF_{t+1}^c.$$

[13]Actually, Barro and Gordon (1983a) postulate a different utility function for the Fed, one that is linear (rather than quadratic) in the output term, and consequently, for them the Ideal Solution is not an achievable outcome. However, they can improve upon the Noncooperative Solution. See also fn. 14.

[14]For simplicity, the Fed is assumed not to discount future utility. If future utility is heavily discounted, two or more reversionary periods might be required; if future utility is very heavily discounted, the approach may not work at all. See Friedman.

[15]To see this, note that $g_t > \pi^* + \delta_t + \bar{\varepsilon} = \pi^* + e_t + \varepsilon_t + \bar{\varepsilon} = g_t^c + \varepsilon_t + \bar{\varepsilon}$ is equivalent to $\varepsilon_t < g_t - g_t^c - \bar{\varepsilon}$.

[16]The reformulation of the model can also be explained in terms of incentive compatibility. Suppose the Fed announces e_t^a and sets $g_t = \pi^* + e_t^a$. Since $g_t^c = \pi^* + e_t$, $g_t = g_t^c$ if and only if $e_t^a = e_t$; no reneging, as defined above, is equivalent to incentive compatibility. Suppose the wage setters revert to inflationary wage settings if $e_t^a > \delta_t + \bar{\varepsilon}$. It is easy to show that $e_t^a > \delta_t + \bar{\varepsilon}$ if and only if $g_t > \pi^* + \delta_t + \bar{\varepsilon}$, and that (22) can be expressed as $P(e_t^a - e_t - \bar{\varepsilon}) = Pr[\varepsilon_t < e_t^a - e_t - \bar{\varepsilon}]$. If the Fed sets $e_t^a > e_t$, it increases the probability of a reversion in period $t+1$. The argument exactly parallels that in the text.

The Fed has no incentive to renege if $U_{g_t}(g_t^c, \bar{\varepsilon}) \leq 0$, where

$$(24) \quad U_{g_t}(g_t^c, \bar{\varepsilon}) = 2y^* - P'(-\bar{\varepsilon})[EUF_{t+1}^c - EUF_{t+1}^{nc}].$$

The first term is the expected gain in the current period from inflating away some of the output differential, θy^*;[17] the second term is the expected loss next period from increasing the probability of a reversion. Note that the increase in the probability of a reversion (evaluated at $g_t = g_t^c$) is just the probability that ε_t equals $\bar{\varepsilon}$;

$$(25) \qquad P'(0-\bar{\varepsilon}) = p(-\bar{\varepsilon}),$$

where $p(\cdot)$ is the probability distribution function of the forecast error, ε_t. From (13) and (14), $EUF_{t+1}^c - EUF_{t+1}^{nc} = y^{*2}/f$; so

$$(26) \quad U_{g_t}(g_t^c, \bar{\varepsilon}) \leq 0 \text{ iff } p(-\bar{\varepsilon}) \geq 2/(y^*/f).$$

If y^*/f is greater than 2 (so that $0 < 2/(y^*/f) < 1$), then $\bar{\varepsilon}$ can be chosen to eliminate any incentive for the Fed to renege.

The trigger mechanism for a reversionary period, $g_t > \pi^* + \delta_t + \bar{\varepsilon}$, was designed to be applicable in the private information case, where wage setters cannot tell whether or not the Fed has reneged. As a result, it has a very interesting property. Even if the Fed is always running the ideal policy, there will still be periodic inflationary episodes associated with large negative prediction errors.[18] Wage setters will see a large g_t and a small δ_t, and they will revert to inflationary wage settings. The frequency of these inflationary reversions depends inversely upon the size of y^*/f. If y^*/f is large, then from (26), a big $\bar{\varepsilon}$ can be specified; consequently, the probability of a reversion is small, and so is the frequency of inflationary episodes.

The modification for private information allows the model to explain periodic inflationary reversions.[19] It also suggests that, despite the apparent instability, Taylor might be right; the Fed and the private sector may have already hit upon a rather good resolution of the problem. In fact, the worse is the problem to begin with, the better is the resolution; θy^* is the difference between the players' output goals, and if y^*/f $(= \theta y^*/(s/\theta))$ is big, a large $\bar{\varepsilon}$ can be specified, and the frequency of inflationary reversions is small. It should also be noted that in this resolution of the credibility problem, the Fed retains full flexibility for stabilization.

There still remain some weaknesses or unanswered questions. First, the game cannot be finite, or as Barro-Gordon have noted, the solution unravels backward. It is not clear how this fact jibes with the finite terms of office of Fed policymakers. Second, it is not clear that the inflationary reversions modeled here correspond to the policy scenario outlined in the introduction. It is tempting to interpret the triggering of reversions as credibility breakdowns, and the inflationary bouts as times of credibility building or learning. But this would not be appropriate. The reversionary periods modeled here are necessary evils that make the Fed's promise of a noninflationary policy credible. Their existence is evidence that the credibility problem of precommitment has been resolved. The instability inherent in a straightforward interpretation of Rogoff's model may more closely correspond to the policy scenario outlined in the introduction.

III. The Legislative Approach with Private Information

The legislative approach might provide a resolution that better fits the game at hand. The social contract enacted can have a finite duration, matching the policymaker's term of office. And a dominant player with some-

[17] It is calculated by substituting $g_t^e = g_t^{e,c} = \pi^*$ into UF_t in (12), differentiating EUF_t with respect to g_t, and evaluating the result at $g_t = g_t^c$.

[18] $g_t^c = \pi^* + e_t > \pi^* + \delta_t + \bar{\varepsilon} = \pi^* + e_t + \varepsilon_t + \bar{\varepsilon}$ is equivalent to $\varepsilon_t < -\bar{\varepsilon}$.

[19] This is the basic insight of Green and Porter.

thing to gain, namely the government, can take the active role in initiating the reform.

Congress could legislate incentive compatible policy rules directly, or it could legislate new rules of the game, rules that would constrain the strategies of the players; targeting procedures are an example of the latter. The solutions described in this section could be legislated either way.[20] However, as was suggested in Section I, it is not straightforward to design incentive compatible rules for monetary policy if the Fed's forecast of money demand is private information. It is easier to motivate the solutions described here in terms of targeting procedures.

The tightest form of money targeting is a k-percent rule requiring g_t to be equal to π^* each period. The noncooperative, *Strict Targeting Solution* is[21]

$$(27)\qquad g_t^{st} = \pi^*, \qquad g_t^{e,st} = \pi^*,$$
$$EUF_t^{st} = -(1+f)\sigma_\delta^2 - y^{*2},$$
$$EUW_t^{st} = -\sigma_\delta^2 < EUW_t^{nc}.$$

Comparing (27) and (13), it turns out that $EUF_t^{st} > EUF_t^{nc}$ if $y^{*2}/f > (1+f)\sigma_e^2$. y^{*2}/f is equal to $EUF_t^c - EUF_t^{nc}$, the social disutility resulting from the noncooperative behavior of the players; $(1+f)\sigma_e^2$ is the social utility derived from having the Fed fully accommodate perceived money demand disturbances.

Legislating a k-percent rule is obviously incentive compatible, and it is better than doing nothing if the disutility of the inflation bias due to noncooperative behavior outweighs the utility of stabilization. Of course, legislating a k-percent rule and doing nothing are the two polar extremes. There are better, less rigid, forms of targeting that limit the Fed's temptation to inflate, but leave it with some latitude for stabilization.

Congress might just require the Fed to meet its target on average, for example. In a two-period social contract, beginning in period $t=1$, *average targeting* requires that

$$(28)\qquad g_1 + g_2 = 2\pi^*.$$

This constrains the Fed's behavior intertemporally; in fact, once g_1 is set, g_2 is known with certainty, by everyone, including the wage setters writing contracts for period $t=2$. The noncooperative, *Average Targeting Solution* is

$$(29)\qquad g_1^{at} = \pi^* + \rho e_1 + y^*/2f,$$
$$g_1^{e,at} = \pi^* + y^*/2f,$$
$$g_2^{at} = \pi^* - \rho e_1 - y^*/2f = g_2^{e,at},$$
$$EUF_t^{at} = -(1+f)\left[\sigma_\delta^2 - (\rho/2)\sigma_e^2\right] - \left[1+(1/4f)\right]y^{*2},$$
$$EUW_t^{at} = -\sigma_\varepsilon^2 - (1/2)\left[1+(1-\rho)^2\right]\sigma_e^2 < EUW_t^{nc},$$

where $\rho = (1+f)/(1+2f) < 1$, and EUF_t^{at} and EUW_t^{at} are expected utilities averaged over the two periods.

Unlike the ideal rule $g_t^c = \pi^* + e_t$, the policy rules in (29) are incentive compatible. Congress could legislate them directly, instead of the targeting procedure; the Fed would announce e_1^a at the time it set g_1, and would have no incentive to make e_t^a greater than e_t.[22]

Here, the Fed will do some stabilizing, but the stabilization will only be partial, and it will only occur in the first period. In addition, the new flexibility results in some inflation bias in the first period; the averaging requires the bias to be reversed in the second period, and this curbs the Fed's temptation to inflate. Compared with strict

[20] They could also be achieved by institutional reform; see fn. 4.

[21] All of the noncooperative solutions in this section are calculated by working backwards, as in Section I.

[22] This can be verified by substituting $g_1 = \pi^* + \rho e_1^a + y^*/2f$ and $g_2 = 2\pi^* - g_1$ into the Fed's two-period optimization problem and maximizing with respect to e_1^a.

targeting, average targeting's advantage is in its flexibility for stabilization; compared with no targeting, average targeting's advantage is in its reduced inflation bias.

In fact, it can be shown that

$$(30)\qquad EUF_t^{at} > EUF_t^{st} \quad \text{if} \quad R > \phi$$
$$EUF_t^{at} > EUF_t^{nc} \quad \text{if} \quad R < (1+\eta)\phi$$

where $R = (1+f)\sigma_e^2/(y^{*2}/f)$

and $\phi = (1+2f)/(1+f)$; $\eta = 2/(1+3f)$.

R measures the utility of stabilization relative to the disutility resulting from noncooperative behavior (i.e., the inflation bias). Note that R depends upon the ability of the Fed to predict money demand disturbances (as measured by σ_e^2), as well as preferences. If $R < \phi$, the Fed's hands should be tied completely with the k-percent rule. If $\phi < R < (1+\eta)\phi$, the Fed should only be constrained by average targeting. If $R > (1+\eta)\phi$, then the Fed's stabilization efforts should not be constrained at all, and the full-inflation bias should be tolerated.

Even here, the full range of possibilities has only begun to be considered. Average targeting is a rather heavy-handed way of achieving flexibility; it is inefficient because it does not link flexibility directly to the size of the disturbances that the Fed ought to be allowed to accommodate. An ideal form of targeting would replace (28) with

$$(31)\qquad g_1 + g_2 = 2\pi^* + e_1 + e_2.$$

However, legislating (31) would not provide an incentive compatible resolution to the credibility problem if the Fed's forecast is private information; the Fed would have an incentive to announce e_1^a and e_2^a larger than e_1 and e_2.[23] Legislating the ideal form of targeting would do no more good than legislating the ideal rule itself.

There is however an incentive compatible form of targeting that is more flexible than average targeting. *Flexible targeting* requires that

$$(32)\qquad g_1 + g_2 = 2\pi^* + c\delta_1, \quad c \geq 0,$$

where δ_1 is of course observed by the wage setters at the end of the first period. With c set equal to one, the noncooperative *Flexible Targeting Solution* $(c=1)$ is

$$(33)\qquad g_1^{ft} = \pi^* + e_1 + y^*/2f,$$
$$g_1^{e,ft} = \pi^* + y^*/2f,$$
$$g_2^{ft} = \pi^* + \varepsilon_1 - y^*/2f = g_2^{e,ft},$$
$$EUF_t^{ft} = \tfrac{1}{2}(1+f)\left(\sigma_\delta^2 + \sigma_e^2\right) - \tfrac{1}{2}f\sigma_\varepsilon^2 - [1+(1/4f)]_y^{*2}$$
$$EUW_t^{ft} = -\tfrac{1}{2}\left(\sigma_\delta^2 + \sigma_\varepsilon^2\right) < EUW_t^{nc},$$

where EUF_t^{ft} and EUW_t^{ft} are expected utilities averaged over the two periods. The policy rules in (33) are incentive compatible and could be legislated directly; the Fed would have no incentive to misrepresent its forecast.[24]

With c set equal to one, the Fed would find it optimal to stabilize fully in the first period. However, it turns out that this is again a polar case. Suppose c is chosen optimally; that is, let c^* be the value of c that maximizes EUF_t^{ft}. It can be shown that

$$(34)\qquad c^* \to 1 \quad \text{as} \quad \sigma_\varepsilon^2/\sigma_e^2 \to 0$$
$$c^* \to 0 \quad \text{as} \quad \sigma_\varepsilon^2/\sigma_e^2 \to \infty$$

and that if c is less than one, the Fed will not find it optimal to fully accommodate

[23] To prove this, first calculate the noncooperative solution with no private information; then replace e_1 and e_2 with e_1^a and e_2^a in the Fed's policy rules for g_1 and g_2, and calculate the Fed's utility-maximizing values for e_1^a and e_2^a.

[24] This can be verified by substituting $g_1 = \pi^* + e_1^a + y^*/2f$ and $g_2 = 2\pi^* + \delta_1 - g_1$ into the Fed's two-period optimization problem and maximizing with respect to e_1^a.

e_1.[25] Full stabilization in the first period should be allowed only if the Fed is very good at forecasting money demand. In fact, with c set equal to one, flexible targeting beats average targeting only if $\sigma_\varepsilon^2/\sigma_e^2$ is less than $(1+f)/(1+2f)$; with c set optimally, flexible targeting always beats average targeting.

It is beyond the scope of this paper to try to identify the most efficient form of targeting. That would require a more interesting control setting (with dynamics, productivity shocks, etc.) and a multiperiod targeting horizon (presumably corresponding to the policymaker's term of office). In addition, Robert Townsend's simple example of multiperiod contracting suggests that a complete specification of the most efficient social contract would be difficult, if not impossible.

The examples provided here do, however, provide two insights that should be helpful in designing efficient targeting procedures. 1) The procedure should be defined in terms of variables that are observed by all market participants; actual money demand was used in the examples above, but actual inflation would have served just as well. Credibility problems will remain if latitude for stabilization is incorporated in any other way. 2) The tightness of the targeting procedure should depend upon the ability of the Fed to stabilize (modeled above as the Fed's ability to predict money demand disturbances that weren't reflected in existing wage settlements) and the importance of stabilization relative to the disutility resulting from the credibility problem (modeled above as an inflation bias).

[25]The general *Flexible Targeting Solution* is

$$g_1^{ft} = \pi^* + [(1+f+cf)/(1+2f)]e_1 + (1/2f)y^*$$

$$g_2^{ft} = 2\pi^* + c\delta_1 - g_1^{ft}$$

$$EUF_i^{ft} = -(1-c)^2[f(1+f)/2(1+2f)]\sigma_e^2 - c^2(f/2)\sigma_\varepsilon^2 - [(1+f)/2](\sigma_\delta^2+\sigma_\varepsilon^2) - [1+(1/4f)]y^{*2}.$$

IV. Conclusion

Kydland and Prescott's game between wage setters and the Fed provides a plausible explanation of the scenario outlined in the introduction, especially if the Fed's forecast of money demand is private information. Without private information, Taylor's observations are persuasive; it is difficult to see why straightforward resolutions of the precommitment problem would not have already been found. But with private information, the Fed cannot demonstrate its forebearance by simply running a noninflationary policy for a few periods, and it is not feasible to simply legislate the "optimal" feedback rule. There are incentive compatible policy rules that could be legislated, but they are more complicated than what is usually envisioned; the rules suggested here can be equivalently encoded as targeting procedures. Private information also enriches Rogoff's model of the perverse policymaker and Barro-Gordon's (1983a) reputation model; both are too stable, without private information, to explain periodic inflationary breakdowns.

There is a recurring tension in this paper that provides a new structure for the old controversy over rules vs. discretion. Some rule or discipline must be placed on the Fed's behavior to resolve the credibility problem and achieve a better, noninflationary, game outcome. The trick is to do this in a way that leaves the Fed with the maximum amount of discretion or latitude for stabilization.[26] The modified Barro-Gordon model resolves the credibility problem in a very efficient manner, as does Rogoff's perversely inflation conscious policymaker (in the absence of productivity disturbances and private information, anyway). In either case, the Fed is left with the latitude to fully accommodate perceived money demand disturbances. The targeting resolutions discussed above do not appear to be as successful; a decrease in the inflation bias is achieved

[26]My use of the word "discretion" differs from that of Barro and Gordon (1983a, b).

at the expense of discretion, the optimal amount of discretion depending upon the ability of the Fed to stabilize and importance of stabilization relative to the elimination of the inflation bias.

More generally, private information should be viewed as a constraint, much like a technology constraint. The job of welfare economists is to develop utility-maximizing mechanisms, given just such constraints. The tradeoff between the discretion needed for stabilization and the rule required for a resolution of the credibility problem is not well understood and deserves more attention.

REFERENCES

Aizenman, Joshua and Frenkel, Jacob, "Wage Indexation and the Optimal Exchange Rate Regime," unpublished manuscript, August 1983.

Baccus, David and Driffill, John, "Rational Expectations and Policy Credibility Following a Change in Regime," *Review of Economic Studies*, January 1985, *52*, 211–21.

Barro, Robert J., "Long-Term Contracting, Sticky Prices, and Monetary Policy," *Journal of Monetary Economics*, July 1977, *3*, 305–16.

______ **and Gordon, David,** (1983a) "Rules, Discretion and Reputation in a Model of Monetary Policy," *Journal of Monetary Economics*, July 1983, *12*, 101–22.

______ **and** ______, (1983b) "A Positive Theory of Monetary Policy in a Natural-Rate Model," *Journal of Political Economy*, August 1983, *91*, 589–610.

Canzoneri, M. B., "The Intermediate Control Problem," *Journal of Money Credit and Banking*, May 1977, *9*, 368–71.

______, "The Time Inconsistency Problem," unpublished manuscript, August 1980.

______ **and Henderson, Dale,** *Strategic Aspects of Policy Making in Open Economies*, forthcoming.

Fischer, Stanley, (1977a) "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economics*, February 1977, *85*, 191–205.

______, (1977b) Long-Term Contracting, Sticky Prices, and Monetary Policy: A Comment," *Journal of Monetary Economics*, July 1977, *3*, 317–23.

______, "Dynamics Inconsistency, Cooperation and the Benevolent Dissembling Government," *Journal of Economic Dynamics and Control*, February 1980, *2*, 93–108.

Friedman, James, "A Noncooperative Equilibrium for Supergames," *Review of Economic Studies*, January 1971, *38*, 861–74.

Gray, JoAnna, "Wage Indexation: A Macroeconomic Approach," *Journal of Monetary Economics*, April 1976, *2*, 221–35.

Green, Edward J. and Porter, Robert H., "Noncooperative Collusion Under Imperfect Price Information," *Econometrica*, January 1984, *52*, 87–100.

Green, S. "Unattainable Objectives, Incomplete Information and Regrettable Macroeconomic Policy," unpublished manuscript, April 1983.

Hart, Oliver, "Optimal Labor Contracts Under Asymmetric Information: An Introduction," *Review of Economic Studies*, January 1983, *50*, 3–36.

Hurwicz, Leonid, "On Informationally Decentralized Systems," in C. B. McGuire and Ray Radner, eds., *Decision and Organization: A Volume in Honor of Jacob Marschak*, Amsterdam: North-Holland, 1972.

Kydland, Finn and Prescott, Edward, "Rules Rather than Discretion: the Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, *85*, 4730–91.

Lee, W., "Optimal Policy Selection and Evaluation with Rational Expectations," unpublished doctoral dissertation, Columbia University, 1981.

Leijonhufvud, Axel, "Constitutional Constraints on the Monetary Powers of Government," presented at the Conference on Constitutional Economics: The Emerging Debate sponsored by The Heritage Foundation, November 1982.

Phelps, Edmund S., "Phillips Curves, Expectations of Inflation and Optimal Employment Over Time," *Economica*, August

1967, *34*, 254–81.

Radner, Roy, "Monitoring Cooperative Agreements in a Repeated Principal-Agent Relationship," *Econometrica*, September 1981, *49*, 1127–48.

Rogoff, Kenneth, "The Optimal Degree of Commitment to an Intermediate Monetary Target," *Quarterly Journal of Economics*, forthcoming.

Roth, Alvin and Schoumaker, Francoise, "Expectations and Reputations in Bargaining: An Experimental Study," *American Economic Review*, June 1983, *73*, 362–72.

Taylor, John, "Comments on 'Rules, Discretion and Reputation in a Model of Monetary Policy'," *Journal of Monetary Economics*, July 1983, *12*, 123–25.

Thompson, Earl, "Who Should Control the Money Supply?," *American Economic Review Proceedings*, May 1981, 356–61.

Townsend, Robert, "Optimal Multiperiod Contracts and the Gain from Enduring Relationships under Private Information," *Journal of Political Economy*, December 1982, *90*, 1166–86.

Waldo, D., "Sticky Nominal Wages and the Optimal Employment Rule," *Journal of Monetary Economics*, May 1981, *7*, 339–53.

The Relevance of Quasi Rationality in Competitive Markets

By THOMAS RUSSELL AND RICHARD THALER*

Smart

My dad gave me one dollar bill
'cause I'm his smartest son,
And I swapped it for two shiny quarters
'cause two is more than one!

And then I took the quarters
And traded them to Lou
For three dimes—I guess he don't know
That three is more than two!

Just then, along came old blind Bates
And just 'cause he can't see
He gave me four nickels for my three dimes
And four is more than three!

And I took the nickels to Hiram Coombs
Down at the seed-feed store,
And the fool gave me five pennies for them,
And five is more than four!

And then I went and showed my dad,
And he got red in the cheeks
And closed his eyes and shook his head—
Too proud of me to speak!

Shel Silverstein
Where the Sidewalk Ends

Economists generally attribute considerable rationality to the agents in their models. The recent popularity of rational expectations models is more an example of a general tendency than a radical departure. Since rationality is simply *assumed*, there is little in the literature to suggest what would happen if some agents were not rational. This is surprising in light of the accumulating evidence that supports Herbert Simon's view that man should be considered at most boundedly rational. In fact, Kenneth Arrow concludes his recent review of this evidence as follows: "I hope to have made a case for the proposition that an important class of intertemporal markets shows systematic deviations from individual rational behavior and that these deviations are consistent with evidence from very different sources..." (1982, p. 8).

In this paper we start to explore the implications of irrationality for economics. We begin by defining what we mean by rational, and what departures from rationality we think are most likely. We then consider what happens if rational and less than fully rational agents (whom we call quasi rational) interact in competitive markets. We show that the knee-jerk reaction of some economists that competition will render irrationality irrelevant is apt only in very special cases, probably rarely observed in the real world. Our analysis highlights the important roles played by arbitrageurs and entrepreneurs. We find that, perhaps counter to intuition, more competition can actually make things worse by leaving no possibility of a profit to an entrepreneur who offers education or information.

I. Rationality, Quasi Rationality, and Framing

Suppose two individuals face the same budget set, but choose different consumption bundles. What could be the reason? Three distinctly different reasons can be identified: 1) the individuals have different tastes (utility functions); 2) the individuals have different information; 3) one of the individuals has made a mistake. In this paper we are primarily concerned with behavior of the third type, so we need a method of modeling mistakes.

There is no place for mistakes in the conventional economics framework. In part, this

*Department of Economics, University of Santa Clara, Santa Clara, CA 95053, and Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853, respectively. The first version of this paper was presented at a conference at Cornell University sponsored by the Center for the Study of the American Political Economy. Thaler wishes to acknowledge and thank the Alfred P. Sloan Foundation for financial support. We have received helpful critical comments from Sherwin Rosen, Joachim Sylvestre, Rex Thompson, and Hal Varian. The usual disclaimer applies.

is because of the difficulty of identifying nonrational (by which we mean non-utility-maximizing) behavior. Consider, for example, the observation of a single purchase. The prices that an agent faces determine a budget hyperplane, and any point on that hyperplane that the agent chooses supports some indifference surface. Thus any chosen point can be consistent with maximization.

Suppose we examine more than one expenditure. Is it possible, on the basis of a *series* of expenditures, to characterize these acts as rational or nonrational? Defining rational now to mean maximizing for a single,[1] increasing concave utility function, the answer is a qualified yes. If the actions contain within them a violation of the weak axiom of revealed preference (which is to say, that we observe both a chosen when b is affordable, and b chosen when a is affordable), then it can be concluded that no fixed increasing strictly concave function is being maximized. Typically economists have looked for violations of the weak axiom in the choices made by agents confronted with different budget sets. Unfortunately, as Hal Varian (1982) has shown, the price income data of the real world seldom oblige by providing other than nested budget sets, so much of the time violations of the weak axiom are simply not possible.

Another way in which, in principle, one could seek violations of the weak axiom is to present the agent with the same budget set, but presented in different ways. Then his choice must not change. This hypothetical test was first suggested by the inventor of the technique of revealed preference, Paul Samuelson (1983). Samuelson considered the following problem. Suppose that we confront an agent with an income-price vector (Y, p) and observe the choice x. Now confront the agent with the income-price vector (mY, mp) where m is a positive constant. Unless the consumer again chooses x, the weak axiom is being violated. The reason is simple. By multiplying both income and prices by m we do not change the budget set. Thus any choice $y \neq x$ at (mY, mp) violates the axiom.

The same approach has been used more recently by cognitive psychologists to demonstrate simple violations of rationality. In a remarkable series of experiments, Daniel Kahneman and Amos Tversky (1979, 1981) have shown that subjects presented with the same problem (budget set) described in different ways repeatedly change their responses. Kahneman and Tversky call such victories of form over substance "framing" effects.

The violations which Kahneman and Tversky find are not only prevalent, they are systematic. That is to say, depending on how the problem is framed, it can be predicted whether the agent will choose x or y. We propose calling any such regular yet nonrational behavior *quasi rational*. Quasi-rational behavior will include some actions that can be considered mistakes in the classification scheme described above. To incorporate such mistakes in a model of a competitive market, an extra feature has to be added to the characterization of consumer behavior. The extra feature captures the consumer's process of translating raw information into a perceived budget set.

Suppose, then, that we think of agents as being given not budget sets but the ingredients from which they can construct a budget set. Call this the information set I. Assume that the individual constructs the budget set B from I using a mapping which we call F so that $B = F(I)$.[2] Once the agent has B, we know from standard duality theory that maximizing choices can be represented as the appropriate derivatives of an indirect utility function U defined on B. With the approach suggested here, U is actually defined on I so that we have $U(F(I))$ as the integral of the maximizing choices.

What is the nature of F? It may have subjective elements but F is not entirely

[1] The word single is crucial, since if tastes are allowed to change, or if a taste for "variety" is permitted, then virtually any set of actions can be rationalized.

[2] The language used here promises more than it delivers. The use of such terms as mapping and information set does not mean that we have a mathematically rigorous theory of this process, and is meant to be suggestive only.

subjective. It should, for example, conform to the laws of mathematics. Thus if I is a sample drawn from some population with replacement, then F should not depend on the order in which the observations are drawn or recorded. In some cases, the mapping from I to B will be so obvious that we would expect no one to get it wrong. Kahneman and Tversky call such mappings *transparent*. The cases of interest are those where the mapping is more complex or, as they say, *opaque*.

There is no shortage of evidence documenting human judgments which fail to satisfy rational objective standards. In many cases (see Kahneman, Paul Slovic, and Tversky, 1982) these lapses seem to be associated with the use of a rule of thumb (i.e., the representativeness heuristic, and the availability heuristic) in which the decision maker sometimes focuses on irrelevant aspects of the information set in constructing his budget set.

This suggests a useful distinction between correct and incorrect mappings. We label the correct (or set of correct)[3] mappings F^* and any incorrect mappings F'.[4] We now have the apparatus to characterize all three reasons referred to earlier why choices (under the same budget set) may diverge: 1) differences in I; 2) differences in U; or 3) differences in F. Those choices consistent with an indirect utility function $U(F^*(I))$ are considered rational while those based on any other mapping, F', are considered quasi rational. The term quasi rational has been chosen to capture both the rational maximizing that is suggested by the systematic regularities shown in the experimental data (subjects do not choose at random) and the inconsistencies with the axioms of rational choice.

The existence of the mapping F is not completely foreign to economics. In expected utility theory, for example, the agent has a preference function over consequences but chooses acts. A mapping F from acts to consequences is needed to construct the budget set. In a subcase of this, portfolio theory of the Markowitz mean-variance type, investors observe the prices of assets but have preferences over the mean and variance of returns. Again a mapping F is needed to construct the relevant budget set. Indeed in financial economics the implication of heterogeneity in F is an ongoing area of investigation (Joram Mayshar, 1983; Robert Jarrow, 1983). We are here assuming not just heterogeneity in F, but the existence of a correct and incorrect F. As we shall see, however, some of the structure is identical. Finally, even in decision making under certainty, the new economic theory of the consumer of Kelvin Lancaster (1966) and Gary Becker (1965) in which consumers have preferences over characteristics but purchase market goods, requires a mapping from goods to characteristics in order to construct the budget set.

The existence of the mapping F enables us to characterize *in principle* even a single choice as nonrational. Suppose we give a consumer information concerning two choices, x and y, and that under the true mapping F^*, x and y are identical but y is cheaper. If an individual buys x, that is nonrational. Nevertheless, it is virtually impossible to classify an act as nonrational *in practice* because of the difficulty in controlling for differences in tastes $U(.)$, or in information I.

This suggests precisely the role of the laboratory experiment. By controlling for tastes and information it is possible to identify F, and in this way conclude that the behavior is indeed nonrational. It is true that for every real world example of quasi-rational behavior we can offer, rationality cannot be ruled out. This, however, in no way rules it in. At the present state of our knowledge, it seems we must allow the possibility that some behavior is quasi rational. Nevertheless, with few exceptions, economists have tended to ignore the work of the cognitive psychologists and have continued to investigate markets with only rational agents. Why

[3]We need not be concerned here with whether F^* is unique. It is sufficient for our analysis that we be able to identify some incorrent mapping F'.

[4]The mixture of subjectivity and required consistency appears also in the subjective theory of probability. It is discussed in very clear terms by Bruno de Finetti (1977).

is the experimental work given so little attention? Economists have generally been critical of these results for the following reasons:

1) "Much of the research is based on hypothetical questions. Thus respondents have little incentive to respond properly." This critique has been examined by David Grether and Charles Plott (1979) and by Grether (1980) in their replications of work by psychologists. In both studies the quasi-rational behavior was at least as strong in a condition with monetary incentives as in a condition with purely hypothetical questions. Thus while skepticism of hypothetical questions may be reasonable, the evidence about quasi rationality cannot be attributed solely to this problem.

2) "The experiments are done in the laboratory." While this statement is true, there is other evidence of irrational behavior outside of the lab. For example, Howard Kunreuther et al. (1978) found irrational factors to be very important in determining who would buy government-subsidized flood insurance.

3) "In the real world, people will learn." There are two responses to this critique. First, the subjects have not yet learned to choose according to our normative theories, otherwise one would not obtain the reported experimental results. Second, as Hillel Einhorn and Robin Hogarth (1978) have emphasized, many situations will not provide feedback in a way that will facilitate learning. Without well-structured feedback, learning may be negligible.

4) "Economists are interested in aggregate behavior and these individual errors will wash out when we aggregate." Since the errors that have been identified are systematic (i.e., in a predictable direction), this statement is just wrong. However, there is a more subtle version of this idea.

5) "Markets will eliminate the errors." While this statement is sometimes made, it is not clear by what mechanism markets will eradicate irrational choices. While it has been argued that evolution will eventually eliminate firms that choose improperly, there is no such process at work for individuals. So far as we know, quasi rationality is rarely fatal.

In summary, there is a large body of experimental evidence suggesting that humans make judgments and decisions in a way that can be characterized as quasi rational. This evidence cannot be dismissed easily. It therefore seems prudent to begin to inquire about the workings of markets in which some agents are quasi rational. Do the quasi rationals affect prices? Does a competitive market protect or exploit the quasi-rational segment? What roles do arbitrageurs and entrepreneurs play? It is to these questions we now turn.

II. Markets with Quasi-Rational Agents

We will investigate competitive markets with two kinds of agents, rational and quasi rational. To capture the quasi-rational behavior, we use an extended Lancasterian model of consumption. Consumers purchase goods in the market but derive utility from the characteristics the goods possess rather than the goods per se. There is an objective mapping from goods to characteristics. The rational consumers perceive this mapping. The quasi rationals perceive a different, incorrect mapping.[5]

A. *Model 1*: *The Basic Model*

We begin by considering the simplest possible model that includes both rational and quasi-rational agents and competition. We make the following assumptions.

Preferences. All individuals are assumed to have the same preferences over two characteristics:

$$U = C_1^{\alpha} C_2^{1-\alpha}. \tag{1}$$

We concentrate our attention on characteristic 1, C_1, so that characteristic 2, C_2, should be thought of as an aggregate of all other characteristics.

Note that if the price of characteristic C_1 is P_{C_1}, the individual demand for characteris-

[5]A similar notion is used in Douglas Auld (1972) and Claude Colantoni et al. (1976).

tic 1, D_{C_1}, is given by

(2) $$D_{C_1} = \alpha Y / P_{C_1},$$

where y is the income of the individual. We assume all individuals have the same income.

The Objective Characteristics Technology. Characteristic 1 is contained in two goods, g_1 and g_2, and only in these goods. Characteristic 2 is contained in the aggregate good g_3 and only in g_3.

By Walras' Law we need only consider equilibria in the markets for g_1 and g_2. We assume

(3) $$g_1 = C_1; \quad g_2 = \beta C_1$$

is the true consumption technology relating quantities of characteristics to quantities of goods.

Quasi-Rational Mapping. Quasi-rational agents believe that the relationship in (3) is actually

(4) $$g_1 = C_1; \quad g_2 = \gamma C_1 \qquad \gamma \neq \beta.$$

Number of Agents. L individuals are rational, M quasi rational.

Supply. There is a fixed supply of goods 1 and 2, $\bar{g}_1, \bar{g}_2$.

Demand. Let P_i = price of good i, $i = 1,2$. If good 1 is bought, the price per unit of characteristic 1 is $P_{C_1} = P_1$. If good 2 is bought, $P_{C_1} = P_2/\beta$. Obviously buyers will buy C_1 at what seems to them to be the lower price. Thus rational demands D^R are given by

$$D_1^R = ((\alpha Y/P_1), 0, (t\alpha Y/P_1))$$
$$\text{as } P_1 < P_2/\beta, P_1 > P_2/\beta, P_1 = P_2/\beta$$
$$D_2^R = ((\alpha Y\beta/P_1), 0, ((1-t)\alpha Y/P_1))$$
$$\text{as } P_1 > P_2/\beta, P_1 < P_2/\beta, P_1 = P_2/\beta,$$

where t is an arbitrary scalar $0 \leqq t \leq 1$. The quasi-rational demands D^{QR} are given by the same expressions with γ replacing β. Note that goods 1 and 2 become perfect substitutes at different prices for the rationals and the quasi rationals.

An equilibrium is defined as a pair P_1^*, P_2^* such that

$$D_1(P_1^*, P_2^*) = \bar{g}_1; \quad D_2(P_1^*, P_2^*) = \bar{g}_2,$$

where D_1 is the aggregate demand for good i.

A Rational Equilibrium is defined as a pair $\hat{P}_1, \hat{P}_2$ such that

$$D_1(\hat{P}_1, \hat{P}_2) = \bar{g}_1; \quad D_2(\hat{P}_1, \hat{P}_2) = \bar{g}_2$$

where $D_i = (L+M)D_i^R$; that is, aggregate demand if all individuals in the economy are rational. We are interested in obtaining necessary conditions for an equilibrium to be a rational equilibrium.

PROPOSITION 1: *A necessary condition for an equilibrium to be a rational equilibrium is that $P_1^* = P_2^*/\beta$.*

PROOF:

Unless the condition is satisfied, rational individuals will not buy both goods. From this we may deduce

PROPOSITION 2: *Let $\gamma > (<)\beta$. Then a necessary condition for an equilibrium to be a rational equilibrium is that $M/L \leq \beta\bar{g}_2/\gamma\bar{g}_1$ $(\leq \bar{g}_1/\bar{g}_2)$.*

PROOF:

Since the proof for $\gamma < \beta$ parallels that for $\gamma > \beta$ we consider only the latter. Take any price ratio $P_1 = P_2/\beta$. Since $\gamma > \beta$, all quasi rationals will buy good 2. Therefore equilibrium requires $P_1 \leqq L\alpha Y/\bar{g}_1$. For rationality we must have $P_1^* = P_2^*/\beta$, so we must have $P_2^* \leqq L\alpha Y\beta/\bar{g}_1$. This, however, would make demand for good 2, $D_2 \geqq M\alpha Y/[\alpha Y\beta/\gamma\bar{g}_1]$. If $\bar{g}_2$ is less than this, that is, $\bar{g}_2 < M\gamma\bar{g}_1/L\beta$, equilibrium is impossible.

Proposition 2 states that a rational equilibrium will not obtain if there are "too

many" quasi-rational consumers. From this it is obvious that

PROPOSITION 3: *There exist equilibria which are not rational equilibria.*

This result demonstrates that the existence of markets is not sufficient to eliminate the effect of quasi-rational behavior. This market, however, has two special features which help sustain the quasi rationality: 1) the only way to trade characteristics is to trade goods; 2) there are no short sales. Both of these features are important and are analyzed in turn.

B. *Markets for Characteristics*

One way in which characteristics could be traded directly is if they could be "stripped" from the goods and sold separately. This may or may not be feasible. For example, it is possible to disassemble an automobile and sell all of its parts, but it is not possible to disassemble and sell its relevant characteristics such as ride, handling, fuel economy, comfort, etc. An extreme case of interest is where characteristics can be decomposed costlessly.

PROPOSITION 4: *If characteristics can be decomposed and marketed costlessly, then a rational equilibrium will obtain.*

PROOF:

Suppose $\gamma > \beta$. Then if equilibrium were not rational, $P_1 < P_2/\beta$. But this would mean that P_{C1} to the quasi rationals who buy good 2 is greater than P_{C1} to the rationals who buy good 1. Thus the law of one price does not hold in characteristics, yielding an incentive for the rationals to buy good 1, strip it of its C_1, repackage C_1 in good 2 and sell it at a profit. (There is likewise an apparent incentive for the quasi rationals to repackage in reverse, but we assume this will immediately reveal the true relationship between goods and characteristics, and so will not occur.) The action of the rationals will thus drive up the price until $P_1 = \beta P_2$.

In goods markets the cost of characteristics stripping is determined by production technology. In asset markets, however, characteristic stripping is not so much a matter of technology as it is the number of markets and cost of using the markets. For example,[6] with a low-cost futures market in Treasury bills, a six-month Treasury bill can be stripped of its three-month component by selling a three-month futures contract in the bill. This means that the price of three-month T-bills and the combination of one six-month T-bill and one three-month future T-bill cannot get far out of line, even if financial officers of corporations have a preference for simple contracts such as three-month bills and are willing to pay a premium for such contracts.

Costless characteristic repackaging de facto sets up a market in characteristics so that the law of one price in characteristics must hold. In the absence of this market, a market in goods is not, in general, a substitute for a market in characteristics.

C. *Short Selling*

In goods markets it is not generally feasible to take a short position. Markets are just not organized in a way that allows a speculator to borrow and sell Chrysler automobiles or Heinz ketchup in the expectation of a future drop in price. Short selling is permitted in some financial markets, however, and so to extend our analysis to include those markets we explore the ramifications of permitting short sales.

PROPOSITION 5: *Short selling will guarantee that the equilibrium is rational provided*: (a) *within the time space that g_1 and g_2 are traded, there exists a time T^* such that after T^* the true relationship between characteristics and goods is known to everyone*; (b) *only the rationals sell short.*

PROOF:

Again assume $\gamma > \beta$. Then if the equilibrium is not rational, $P_2 > P_1\beta$. But at time T^*, $P_2 = P_1\beta$. Thus short selling by the rationals will be profitable and will force P_2 into equality with $P_1\beta$.

[6]This example is discussed in Edwin Elton et al. (1982).

The two extra conditions are necessary to assure that an equilibrium exists and that short selling is riskless.[7] If the quasi rationals sell short, and if $P_2 = \beta P_1$, they will believe g_1 is overvalued and will wish to sell it short. Thus no equilibrium will exist. The assumption that only rationals sell short is not unreasonable if rationality is associated with professional market participants. It might be called the Marshallian view based on the following from Alfred Marshall:

> The private purchaser of railway shares may know nothing [about its prospects, the ability of its management and propriety of its accounts], but he buys with the confidence that all such points have been scrutinized by many keen men with special knowledge, who are able and ready to "bear" the stock if they find in it any weak spot, which...had not been allowed for in making up its value.
>
> [cited in Mayshar, pp. 126–27, fn. 25]

The condition that the true mapping be revealed is necessary to create a pure arbitrage opportunity (some chance of gain, no chance of loss). If characteristic stripping is impossible, then knowledge of the true mapping between goods and characteristics is not sufficient to create an arbitrage opportunity. Only if the quasi rationals become informed can the correct price be assured.

Of course, in most nonfinancial markets, characteristic stripping is not costless and short selling is impossible. In these situations (within Model 1) quasi rationals do influence prices, and the rational price equilibrium need not obtain.

D. *Production*

Up to now we have been assuming that both goods are in fixed supply. To allow for production we will consider three cases: increasing costs, constant costs, and decreasing costs. Formal proofs follow the same lines as above so we just present the results.

Increasing Costs. If both goods are produced with technologies involving increasing marginal costs, then the results of the previous section are qualitatively the same. A rational equilibrium can obtain as long as the number of quasi rationals in the economy is small enough.

Constant Marginal Costs. The constant marginal cost case is a knife-edge situation. Competition assures that price equals marginal cost so both goods can only coexist if the ratio of marginal costs is exactly equal to β. The size of the two groups of consumers is irrelevant.

Decreasing Marginal Cost. When both goods are produced with economies of scale, then a rational equilibrium can occur with both goods existing. Also, if the quasi rationals are large enough in number and the goods are close enough in efficiency cost, then the quasi rationals can lead to the wrong good being produced. The rationals in this case recognize that it will be cheaper for them to join the quasi rationals than to buy their preferred (*ex ante*) good.[8]

E. *Comparison with Results in Finance*

The models of fixed supply presented so far are very close in structure to a class of models in financial economics first introduced by John Lintner (1969).

In these models, individuals have different beliefs concerning the mean and variance of assets. The reasons for these different beliefs are not investigated so there is no counterpart to our notion of rationality and quasi rationality, but still the market is composed of individuals with different beliefs, and this assumption might be expected to produce similar results.

In an important sense, however, our results are quite different. In the financial mod-

[7] In the finance literature on arbitrage pricing, these two conditions appear as: 1) all investors agree on the state representation; 2) all investors agree on probability zero events.

[8] A similar problem is analyzed by John Haltiwanger and Michael Waldman (1985). They call increasing costs congestion and decreasing costs synergy.

els à la Lintner, market prices reflect *all* beliefs. For that reason Lintner himself found the extension of the model to heterogeneous beliefs basically uninteresting, since everything that was true for homogeneous individuals now became true for "the market." In the models we discuss, market prices may give zero weight to the beliefs of one class of agents. Why the difference? The reason hinges on special assumptions made in the financial literature which prevent the financial system from breaking up into subsystems. Because assets are assumed to be joint normally distributed and agents are assumed to have exponential utility functions, all individuals hold all assets. For that reason, asset prices reflect all beliefs. In our model, because there are more goods than characteristics and because the technology is linear, it is possible for rational agents completely to escape the influence of quasi-rational agents by specializing in consumption of the good(s) which the quasi rationals cause to be underpriced. It is precisely this force that can restore rationality to the market. Obviously a necessary condition for this to occur is that there are more goods than characteristics. As the finance examples make clear, this is not sufficient. If, for any reason, all goods are bought by all agents, quasi rationals must influence prices. Thus the examples we discussed earlier are actually biased towards the result that market prices will be fully rational, since they permit rational agents to form their own subeconomies.

III. An Example

As we emphasized in the introduction, it is generally not possible to prove that any act or set of actions is generated by quasi-rational behavior. Differences in tastes and/or information can rarely be ruled out completely. Nevertheless, we present here some data from a market where the law of one price (for characteristics) is violated. While a plausible quasi-rational explanation can be given, as usual rational-based explanations can also be made. Our purposes, therefore, are just to give an example of how a market might turn out when a quasi-rational rule of thumb is widely used, and to use the example to address some other theoretical issues.

A. *Dishwashing Liquid*

In 1981 Consumers Union (CU) conducted a study of the price and efficiency of dishwashing liquids. The study was replicated in 1984 with very similar results. We will present the more recent data. Thirty-five brands of dishwashing liquid were tested for their ability to wash dishes. Few differences among brands were discovered on most dimensions, but wide variation was found on the number of dishes a given quantity of the brand could wash. This "dishes washed per squirt" measure was called an efficiency factor. Brands were placed into four groups according to their efficiency factor. Brands in the top group were arbitrarily given an efficiency factor of 1.0. By multiplying the nominal price of the brand by the efficiency factor, a "real cost" was calculated. If the law of one price holds for the characteristic dishes washed, then the real cost of each brand should be about the same.

Table 1 presents CU's results. As can be seen, the law of one price fails to hold. There is a clear negative relationship between the nominal selling price and the real cost. The most expensive brands are usually the best buys. There may, of course, be other characteristics, such as kindness to hands. However, CU found little difference on these dimensions. Furthermore, the most expensive brands are likely to have more of *all* the (positive) characteristics so adding more characteristics would probably strengthen the results.

We think the most plausible explanation for this finding is that some consumers confuse the mapping from price per bottle to price per dish washed. It is well known in marketing that many consumers have a general tendency to buy either the cheapest brand or the most expensive brand. This tendency represents a shopping strategy that greatly reduces decision-making costs at the supermarket. It may well be *rational* to use such a strategy. It would take so long to fully analyze every decision for a single week's family shopping expedition that some sim-

TABLE 1—REAL COSTS OF DISHWASHING LIQUID

Group	Number of Brands in the Group	Average Price	Efficiency Factor	Real Cost
A	8	1.97	1.0	1.97
B	6	1.60	1.4	2.24
C	19	.97	2.7	2.61
D	2	.72	4.7	3.36

Source: Consumer Reports, July 1984, p. 413.

plifying strategies must be used. In cases where quality or taste is easy to judge, a family may learn to make specific alterations to their general strategy ("Don't buy generic cola"). In other cases, such as the dishwashing liquid, a family would have to do some fairly sophisticated testing to determine that its usual "buy cheap" strategy was (in this case) inappropriate.

Since this market has remained stable over the last few years (and probably for much longer), it becomes interesting to ask why the inefficient brands survive. We will consider four forces that could push the (characteristics) market back into equilibrium: arbitrage, entry, education, and tied sales.

Arbitrage and Entry. Arbitrage would be possible if one could profitably buy the expensive brands, dilute them, and sell the diluted product as a cheap brand. However, there is no reason to think this is possible. Entry into the "no frill," generic dishwashing liquid market is relatively free. There are unlikely to be profits to be made by entering this market. The high real cost of these brands probably represents the high fixed cost of packaging and distribution. Literally buying the high-priced brands off the shelf and diluting them for resale is surely an unprofitable venture, and since the data do not necessarily imply extraordinary profits in any segment, entry alone cannot be expected to solve the problem.

Education. One of the high-price/low-cost brands of dishwashing liquid has, from time to time, run an advertising campaign that stresses the true economy of their brand relative to the low-cost "so-called bargain brands." This is an example of a firm trying to educate the quasi-rational segment. Whenever a consumer can be educated at a cost that is less than his potential gain from switching to the efficient product, a profit opportunity exists. However, ironically, this education will not take place if the market is truly competitive. With perfect competition, no one seller can charge a premium above marginal cost and so there is no incentive to pay the costs of education. Only if there is some monopolistic element, such as brand names, will there be a potential return to education. Even then, the educator runs the risk that the education will not be brand specific, so other high-cost brands may be able to free ride at the educator's expense.

Tied Sales. Jerry Hausman (1979) has done a careful study of consumer purchases of air conditioners. He finds, much as in the dishwashing liquid example, that more expensive air conditioners are better buys because they are generally more energy efficient. He reports that the average purchase implies a discount rate of 25 percent after considering utilization rates and energy costs. Furthermore, the implicit discount rates vary systematically with income. Purchases by low-income households imply discount rates of 27, 39, and 89 percent in the three lowest income classes in Hausman's sample. These rates are all much higher than the prevailing borrowing rates (around 18 percent on most credit purchases) at that time.[9]

Hausman discusses several possible solutions to the apparently inefficient purchases being made. One is of particular interest here:

> Another possible type of market solution would be to have utility companies purchase appliances and lease them to their customers. Presumably utilities would be willing to engage in such activity, since they could borrow money to finance the more energy-efficient

[9]Air conditioners are rarely purchased by the very poor so most buyers probably have access to at least installment-buying-type credit.

> appliances and then charge a rental rate which would leave the consumer better off. Utilities could develop expertise in choosing the optimal efficiency model in terms of climate and intended utilization and help their customers make a better choice. [p. 51]

While Hausman's idea is along the right lines, it may not go far enough. What the utility would have to do to be sure of getting optimal choices is to rent the air conditioners with the utility costs included in the rental. Only by tying the sale of the air conditioner services with the purchase of the electricity could the possibility of quasi-rational choices be ruled out. Of course, other problems such as monitoring utilization might prevent such an arrangement from succeeding. Nevertheless, the theoretical point of interest here is that only by creating a market in the ultimate consumption commodity (the characteristics in the model) can the seller guarantee rational choices.

It is interesting to compare this conclusion with that made by Richard Posner in a similar case:

> The leverage theory (of tie-in sales) held that if a seller had a monopoly of one product, he could and would monopolize its indispensable complements as well, so as to get additional monopoly profits. Thus, if he had a patented mimeograph machine, he would lease the machine at a monopoly price and also require his lessees to buy the ink used in the machine from him and charge them a monopoly price for the ink. This procedure, however, makes no sense as a matter of economic theory. The purchaser is buying a service, mimeographing. *The pricing of its components is a mere detail*; it is, rather, the total price of the service that he cares about. If the seller raises the price of one component, the ink, the purchaser will treat this as an increase in the price of the service. If the machine is already being priced at the optimal monopoly level, an increase in the price of the ink above the competitive level will raise the total price of the service to the consumer above the optimal monopoly level and will thereby reduce the monopolist's profits.
> [1979, p. 929, emphasis added]

Posner, of course, explicitly assumes rational consumers. He says that to do otherwise would be "inconsistent with the premises of price theory." But if even some consumers are quasi rational, then the way the prices of the various components of a good are framed is no longer a "mere detail." Indeed, framing effects in particular, and quasi rationality generally, open the possibility that repackaging goods via tie-in sales and other similar devices can increase both consumer welfare and monopoly profits. Thus the "Chicago" position on tie-in sales (to permit them) may be right, though for the wrong reasons.

V. Conclusion

> When we assume that consumers, acting with mathematical consistency, maximize utility, therefore, it is not proper to complain that men are much more complicated and diverse than that. So they are, but if this assumption yields a theory of behavior which agrees tolerably well with the facts, it must be used until a better theory comes along.
> [George Stigler, 1966, p. 6]

There are two possible justifications for the use of maximizing models in applied microeconomics. As Stigler suggests above, one justification is that the models are good predictors. This is the usual "as if" position. The alternative justification is that markets guarantee that only rational behavior can survive. Our reading of the psychology literature referred to earlier suggests that the first justification is frequently violated. Deviations from maximizing behavior are both common and systematic. The implication of the current paper is that the second justification will rarely apply, except (perhaps) in some highly efficient financial markets. Where does that leave us?

First of all, our analysis suggests that research on individual decision making is

highly relevant to economics whenever predicting (rather than prescribing) behavior is the goal. The notion that individual irrationalities will disappear in the aggregate must be rejected. However, as Stigler implies, the neoclassical theory will not be abandoned until an acceptable (superior) alternative is available. Such theories will have to be explicitly descriptive rather than normative. The usual approach in economics is to solve for the optimal solution to a problem, and then to *assume* the agents in the model chose accordingly. Thus the model is supposed to be simultaneously normative and descriptive. A model such as Kahneman and Tversky's (1979) prospect theory abandons any claim to normative value. It simply tries to describe human behavior. Thaler and H. M. Shefrin's 1981 self-control theory of saving is in a similar spirit. Both theories seem to fit the data well. It is worth mentioning that both of these models are still basically maximizing. Quasi rationality does not imply random choice.

In the absence of such behaviorally motivated alternative theories, one intermediate step can be taken. A standard practice in applied work is to use the theory to impose restrictions to the empirical estimates. The estimates are then forced to satisfy the restrictions. In the absence of evidence to support the *assumption* that the theory describes behavior, a simple precaution is to do the estimates in an unconstrained fashion whenever that is possible. For example, Grether gave subjects in an experiment a Bayesian revision task in which they *should* equally weight the (given) prior odds and likelihood ratio. He then estimated how they *did* combine the data and found that the subjects on average overweighted the likelihood ratio. The model he estimated would outperform an alternative model that assumed proper Bayesian revision. Until better theories are developed, such atheoretical estimation procedures seem appropriate.

Our analysis also has implications for the use of evolutionary arguments in economics. In a review of Richard Nelson and Sidney Winter's 1982 book on this subject, Michael Spence says that "markets discipline agents and modify their behavior." This statement is clearly true for agents within firms, but has limited applicability to individuals acting as consumers or investors. In fact, the more efficient the market, the *less* discipline the market provides. In a fully arbitraged market, all goods (assets) yield the same characteristics per dollar (returns), thus individuals can choose in any manner without penalty. Only in less than fully efficient markets is there any penalty to quasi rationality.

REFERENCES

Arrow, Kenneth, "Risk Perception in Psychology and Economics," *Economic Inquiry*, January 1982, *20*, 1–9.

Auld, Douglas, "Imperfect Knowledge and the New Theory of Demand," *Journal of Political Economy*, November/December 1972, *80*, 1287–94.

Becker, Gary S., "A Theory of the Allocation of Time," *Economic Journal*, September 1965, *75*, 493–517.

Colantoni, Claude S., Davis, Otto A. and Swaminuthan, Malati, "Imperfect Consumers and Welfare Comparisons of Policies Concerning Information and Regulation," *Bell Journal of Economics*, Autumn 1976, *7*, 602–18.

Einhorn, Hillel J. and Hogarth, Robin M., "Confidence in Judgement: Persistence in the Illusion of Validity," *Psychological Review*, No. 5, 1978, *85*, 395–416.

Elton, E., Gruber M. and Rentzler, J., "Intra Day Tests of the Efficiency of the Treasury Bills Futures Market," Working Paper No. CSFM-38, Columbia Business School, October 1982.

de Finetti, Bruno, *Theory of Probability*, Vols. 1; 2, London: Longmans, 1977.

Grether, David, "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *Quarterly Journal of Economics*, November 1980, *95*, 537–57.

_____ **and Plott, Charles,** "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review*, September 1979, *69*, 623–38.

Haltiwanger, John and Waldman, Michael, "Ra-

tional Expectations and the Limits of Rationality: An Analysis of Heterogeneity," *American Economic Review*, June 1985, *75*, 326–40.

Hausman, Jerry, "Individual Discount Rates and the Purchase and Utilization of Energy-Using Durables," *Bell Journal of Economics*, Spring 1979, *10*, 33–54.

Jarrow, Robert, "Beliefs, Information, Martingales and Arbitrage Pricing," Working Paper, Cornell Graduate School of Management, 1983.

Kahneman, Daniel, Slovic, Paul and Tversky, Amos, *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press, 1982.

______ **and Tversky, Amos,** "The Framing of Decisions and the Psychology of Choice," *Science*, January 1981, *211*, 453–58.

______ **and** ______, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, March 1979, *47*, 263–91.

Kunreuther, Howard et al., *Disaster Insurance Protection: Public Policy Lessons*, New York: Wiley & Sons, 1978.

Lancaster, Kelvin J., "A New Approach to Consumer Theory," *Journal of Political Economy*, April 1966, *74*, 132–57.

Lichtenstein, Sarah and Slovic, Paul, "Reversal of Preferences Between Bids and Choices in Gambling Decisions," *Journal of Experimental Psychology*, 1971, *89*, 46–55.

Lintner, John, "The Aggregation of Investors' Diverse Judgments and Preferences in Purely Competitive Markets," *Journal of Financial and Quantitative Analysis*, December 1969, *4*, 347–400.

Mayshar, Joram, "On Divergence of Opinion and Imperfections in Capital Markets," *American Economic Review*, March 1983, *73*, 114–28.

Nelson, Richard and Winter, Sidney, *An Evolutionary Theory of Economic Change*, Cambridge: Harvard University Press, 1982.

Posner, Richard, "The Chicago School of Antitrust Analysis," University of Pennsylvania *Law Review*, 1979, *127*, 925–52.

Samuelson, Paul A., *Foundations of Economic Analysis*, Cambridge: Harvard University Press, 1983.

Shefrin, H. M. and Thaler, Richard, "Life-Cycle vs. Self-Control Theories of Saving: A Look at the Evidence," unpublished paper, Cornell University Graduate School of Management, 1984.

Stigler, George, *The Theory of Price*, New York: Macmillan, 1966.

Thaler, Richard, "Toward a Positive Theory of Consumer Choice," *Journal of Economic Behavior and Organization*, 1980, *1*, 39–60.

______ **and Shefrin, H. M.,** "An Economic Theory of Self-Control," *Journal of Political Economy*, April 1981, *89*, 201–2.

Varian, Hal R., "The Nonparametric Approach to Demand Analysis," *Econometrica*, July 1982, *50*, 945–73.

Efficient Adaptation in Long-Term Contracts: Take-or-Pay Provisions for Natural Gas

By SCOTT E. MASTEN AND KEITH J. CROCKER*

To avoid repeated bargaining in transactions supported by durable, transaction-specific investments, parties may decide to specify the terms of future trade in a long-term contract at the outset of the relationship (see, for example, Benjamin Klein, Robert Crawford, and Armen Alchian, 1978; and Oliver Williamson, 1979). A principal limitation of long-term contracting, however, is its inflexibility in the face of fluctuations in supply and demand: although contingent claims contracts permit adaptation to changing circumstances, contingent performance is costly to stipulate and even more difficult for courts to administer. To mitigate these hazards, parties will therefore wish to choose contract terms that minimize the need for costly adjudication while maintaining incentives for appropriate adaptation.

This paper examines the incidence of "take-or-pay" provisions in contracts between natural gas producers and pipelines from this perspective. Take-or-pay clauses require purchasers to pay for a contractually specified minimum quantity of output, even if delivery is not taken. The existence of such provisions and efforts by pipelines to have them abrogated in the face of declining demand during the most recent recession have generated a debate regarding the role such terms play in allocating gas resources. A common perception has been that take-or-pay provisions are an artifact of wellhead price regulation: by assuring a minimum payment, take obligations raise the expected value of a contract to a producer and thereby circumvent the effect of the price ceiling. To the extent that producers of high-cost gas were able to command higher obligations, some pipelines have been induced to purchase and sell more expensive gas to end users while leaving lower-cost supplies with smaller take requirements in the ground. The implication of this view is that take provisions are anomalies that distort market incentives and should therefore be nullified, permitting pipelines to adjust purchases in a more appropriate manner.

The problem with this explanation is that the incidence of take provisions is not limited to regulated gas supplies, but is also a feature of contracts covering unregulated pre-1954 interstate and pre-1978 intrastate gas, as well as of recently deregulated categories of new "high-cost" gas.[1] Moreover, take-or-pay clauses are also encountered in contracts for

*Graduate School of Business Administration, University of Michigan, Ann Arbor, MI 48109, and Department of Economics, University of Virginia, Charlottesville, VA 22901, respectively. This paper was presented at the Third Annual Conference of the Rutgers University Advanced Workshop in Public Utility Economics and Regulation, May 1984. We thank Michael Canes, Maxim Engers, Matt Gelfand, Richard Higgins, Douglas Kinney, Jonathan Skinner, the University of Virginia Microeconomics Workshop, and an anonymous referee for helpful suggestions. Financial support was provided by the Federal Trade Commission. The views presented are our own, not those of the Commission, individual commissioners, or staff.

[1] See H. G. Broadman and M. A. Toman (1983); and M. E. Canes and D. A. Norman (1983). Take provisions were also an integral part of producer-pipeline contracts throughout the essentially unregulated period before area pricing was adopted in 1960. The *Phillips* decision in 1954 gave the Federal Power Commission jurisdiction to regulate wellhead prices of gas sold in interstate commerce. The initial regulatory efforts, however, sought to institute cost-based rates on a well-by-well basis. Using this approach, the Commission never progressed beyond regulating the parties to the original decision. The result was a prodigious case backlog. According to Paul MacAvoy and Robert Pindyck, "...The Commission itself forecast that it would not finish its 1960 case load until the year 2043" (1975, p. 13). Area rates, that effectively froze prices at the market levels of 1958–59, were an attempt to lend tractability to the regulatory process.

coal and other unregulated commodities.[2] In view of this, both H. G. Broadman and M. A. Toman, and M. E. Canes and D. A. Norman have suggested that take obligations might be a means of allocating risk between producers and pipelines. High take provisions reduce risk for producers by guaranteeing a minimum return on investments in well capacity. Unfortunately, such arguments do not provide a practical basis upon which to evaluate observed contractual arrangements without knowledge of the relative risk preferences of the parties involved.

This paper offers an interpretation of take-or-pay provisions that relies on neither risk aversion nor the existence of regulatory price ceilings. Instead, we argue that take obligations can be viewed as a mechanism for effecting appropriate incentives for contractual performance, and show that efficient breach considerations define an optimal take percentage as a function of characteristics of the transaction. These incentives are distorted, however, by the existence of regulated price ceilings, causing the adoption of take obligations *in excess* of optimal levels.[3] Whether a policy to reduce "excessive" take provisions to *ex post* optimal levels can be justified on the basis of regulatory interference depends on the regulatory environment expected to govern the development of future gas reserves.

Section I develops these arguments and applies them to contracting in the natural gas industry. Section II presents an empirical test of the model employing actual data on producer-pipeline contracts and well characteristics.

I. Performance Incentives in Long-Term Contracts

Once a transaction-specific investment has been made, only imperfect market alternatives exist and both the buyer and seller are locked into a bilateral monopoly relationship. To prevent contention over the resulting quasi rents from dissipating too large a portion of the gains from trade, the parties may try to secure a mutually advantageous distribution through a contract, the duration of which will depend in part on the durability of the associated investments. In industries with particularly durable capital, it is not uncommon to observe contractual agreements that extend for ten years or more.[4]

Over such long horizons, the need for adaptation to changing circumstances, and hence the desire for flexible arrangements, may be substantial. But a tradeoff generally exists between the flexibility provided for in a contract and the ease with which it can be implemented: a single contractual stipulation is relatively straightforward for courts to enforce in comparison to multiple contingent claims which require that both the parties and the courts establish the state that has actually transpired. The more provisions stipulated, the greater the scope for both honest misinterpretation and intentional deception, and thus the greater the likelihood of a dispute requiring costly adjudication.

To minimize these costs, the parties will wish to stipulate terms that do not require court verification of exogenous events. Accordingly, contracts usually employ unilateral options rather than contingent clauses to accommodate adaptation.[5] The goal is to

[2]As Canes and Norman note, E. M. Carney (1978) discusses the use of take-or-pay provisions in coal contracts. Similar arrangements also appear in other contracts as minimum bill provisions.

[3]Except where explicitly noted, we will be referring throughout the paper to private optimality in exchange between a buyer and a seller. The social efficiency of any contractual provision depends upon whether private valuations are distorted from social valuations by the existence of externalities or other factors. We show here that the presence of price regulation does not in and of itself justify the abrogation of take-or-pay obligations.

[4]Victor Goldberg and John Erickson, for instance, note that in their sample, "Nine of ten contracts...involving new [petroleum] cokers were for a period of at least ten years" (1982, p. 10). Coal contracts are of similar duration (Carney, p. 197). Also see below.

[5]The notion of a unilateral option is analogous to the self-selection behavior often observed in theoretical models of bargaining in an environment of asymmetric information (see, for example, Milton Harris and Robert Townsend, 1981).

design contracts in ways that reconcile the exercise of such options with joint profit-maximizing behavior.[6]

The adoption of take-or-pay provisions in long-term contracts can be usefully interpreted in this light. Fluctuations in demand or costs may make it unprofitable or even inefficient to carry out the original objectives of a contract. By altering incentives to accept or reject delivery, take provisions can induce buyers to release investments to their alternative uses only when it is efficient to do so.

A. *Natural Gas Production and Contracting*

These considerations bear directly on the organization of production and exchange in the natural gas industry. Most natural gas is purchased by pipelines from independent producers for distribution to customers in regional markets. Like oil production, the extraction of gas requires large, durable, location-specific investments in facilities and equipment. However, unlike the field market for oil which is characterized by a functioning spot market, gas sales tend to be governed by extended contracts averaging fifteen to twenty years in length. This disparity in contractual structures governing commodities that are so closely related in production technologies may be traced to differences in transmission alternatives. Whereas pipelines represent virtually the only economically feasible form of transportation for gas, oil may be transported by truck or barge as well as by pipeline, thereby reducing the extent to which oil producers are locked into a bilateral relationship.[7]

Uncertainty regarding future market conditions can make it hazardous to commit resources contractually to a particular application, however. For example, in the face of a decline in gas demand, it may become more economic to sell gas to an alternative pipeline (if the decline is regional) or to store it for future use (if the decline is economy-wide). Since uncertainty increases with the distance of the relevant horizon, the need to provide for adaptation is most acute in longer term agreements. The prevalence of take-or-pay clauses in both long-term coal and gas contracts is consistent with this reasoning. Take obligations encourage efficient adaptation by relating the payment schedule in a contract to the alternative values of the resources either in sale to alternative customers or in storage for future use.

Alternative sale values of a product in transaction-specific relationships are limited by the design and location of specialized investments. For natural gas, the most important determinant of that value is the number and proximity of alternative pipelines. The fewer the connections to pipelines, the less likely that a producer will be able to dispose of gas at a price comparable to that in the original contract. In the extreme, the sale of gas to an alternative customer may require the construction of costly new transmission facilities.[8]

[6] That joint profit-maximizing behavior does not always coincide with the private incentives of the parties to an exchange is a familiar proposition:

> ...joined as they are in an idiosyncratic condition of bilateral monopoly, both the buyer and seller are strategically situated to bargain over the disposition of any incremental gain whenever a proposal to adapt is made by the other party. Although both have a long-term interest in effecting adaptations of a joint profit-maximizing kind, each also has an interest in appropriating as much of the gain as he can on each occasion to adapt. Efficient adaptations which would otherwise be made thus result in costly haggling or even go unmentioned, lest the gains be dissipated in costly subgoal pursuit. [Williamson, p. 242]

[7] A more complete discussion of oil field markets and how they differ from those of gas is given in S. L. McDonald (1971).

[8] Alternative sale possibilities are sometimes governed by the particulars of the supply contract between a producer and a pipeline. Often, the contract governs only a portion of the well's output, which is sold to the contracting pipeline on a first refusal basis. In other cases, particularly, when the field is served by a single pipeline, the entire output of a well is "dedicated" to a particular pipeline for the life of the contract. When gas demand is low, however, pipelines have little incentive to enforce these dedication clauses. Indeed, given the present surplus, many pipelines are attempting to renegotiate their purchase obligations downward and would welcome sale to alternative buyers (see D. Norman, 1984).

Because of this, and the fact that the demand for energy resources often fluctuate on a nationwide basis, the best alternative employment for both coal and gas is often to store it for future use. Since the product can usually be left in the field, storage is generally less costly for the producer than the customer. The resource value of stored gas may be diminished, however, by the proximity of other wells: gas not extracted may be drained away by other producers operating in the same field. To attenuate the problem of competitive extraction and protect landowners from drainage, the primary gas-producing states have instituted well-spacing rules and prorationing of output among the wells in a field.[9] The prorationing formulae generally used assign each producer an interest in the field based on surface ownership and well deliverability. Gas demand is divided among the various producers by the assignment of production allowables based on each producer's interest in the field.

But such intervention, while mitigating the drainage problem, does not eliminate it. The formulae are, at best, imprecise rule-of-thumb estimates of gas location relative to surface ownership. Moreover, a production allowable conveys only the *opportunity*, not the guarantee, to produce a given amount of gas in a specified period of time. An operator who is unwilling or unable to produce his allowable faces the prospect of drainage to those who do.[10] Consequently, most gas production remains governed by the *Rule of Capture*: "Possession of the land...is not necessarily possession of the gas. If an adjoining, or even distant, owner drills on his own land, and taps your gas, so that it comes into his well and under his control, it is no longer yours, but his."[11]

In general, the desirability of reallocating resources away from their intended use depends critically on their value in alternative applications. In the case of gas, the appropriate take obligation in each instance will depend on the nature of the well and its relation to other wells, piplines, and markets.

B. *Take-or-Pay and Breach of Contract*

To illustrate these concerns, consider the relationship between a pipeline (or buyer) and producer (or seller) of natural gas, both of whom are assumed to be risk neutral. After gas has been discovered but prior to investing in production and transmission facilities, the parties write a contract specifying a capacity level and the terms under which the product is to be exchanged in subsequent periods. The value of this capacity to the pipeline depends upon such things as weather patterns, economic fluctuations, and the prices of alternate fuels, all of which are uncertain at the time the contract is written and the well drilled. If we let θ represent this uncertainty, and define

$v(\theta)$ = the value of the well to the pipeline, net of transmission costs and gross of payments to the producer; and

y = the payment made by the pipeline to the producer for a contractually specified quantity of gas;

then, once θ is revealed, the pipeline and the producer would receive $v(\theta)-y$ and y, respectively—*if* the exchange takes place: for low values of θ, and hence of v, the pipeline would wish to breach the contract with the producer. Specifically, the pipeline would wish to discontinue deliveries whenever $v(\theta) < y$.[12] Were this to occur, the pipeline would

[9] Gas and oil prorationing differ in several significant respects, reflecting the greater importance of extraction rates on the amount of oil that is ultimately recoverable. For a more complete discussion, see R. E. Sullivan (1955).

[10] Both Texas and Oklahoma have attempted to ensure that producers are equally able to sell their allowables by requiring pipelines to purchase ratably (according to each producer's interest) from their suppliers in a field. However, other states do not require ratable take (Sullivan, p. 348), nor does this approach preclude disproportionate purchases by different pipelines.

[11] *Westmoreland and Cambria Natural Gas Company v. Dewitt, 130 Pa. St. 235, 18A. 724, 725* (1889).

[12] For analytical tractability, we consider a model in which gas deliveries are discontinued in a discrete fashion; i.e., the purchaser either "takes" the gas at the contract price, or else "pays" the contractually specified percentage of the full obligation and releases all of the gas contracted for to its alternative use. With a few exceptions, the continuous analog to this model in which

then earn no net return,[13] and the producer would seek the next highest value of his capacity, $s(\alpha)$, which is a function of well attributes, α.

Since efficiency requires that gas be used in its highest value, breach would in fact be efficient if $s(\alpha) > v(\theta)$. But for $s(\alpha) < v(\theta) < y$, the pipeline would wish to breach even though it would be inefficient to do so. Figure 1 depicts the ranges of $v(\theta)$ for which the buyer would wish to breach and for which breach would be efficient. In general, there is a tendency, as illustrated in the diagram, for breach to occur too frequently in unsecured agreements.

The pipeline could, however, be induced to reject delivery efficiently by imposing a penalty for nonperformance, $\delta = y - s(\alpha)$. This penalty, known as "expectation damages," is commonly employed by the courts and would normally apply to gas contracts in the absence of a stipulated take obligation.[14] If the courts always and unerringly chose this award, there would be little need to stipulate damages in the contract. But the uncertainty associated with judicial rulings encourages costly litigation. By stipulating damages, the parties avoid the costly process of determining the appropriate penalty in the courts, and rejection of deliveries becomes an option that may be unilaterally invoked by the purchaser.

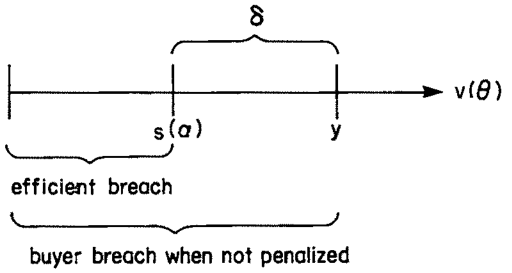


FIGURE 1

Since the optimal penalty for breach declines as a well is depleted, gas contracts usually express the penalty for refusing delivery as a fraction of reserves or deliverable capacity. In that way, the penalty obtaining in each successive period covered by the contract adjusts automatically to the declining level of remaining reserves. Written as a percentage, γ, of the contractually specified payment, the penalty described above becomes

$$(1) \qquad \gamma = 1 - s(\alpha)/y.$$

Since y must be at least as great as $s(\alpha)$ to cover the fixed costs of production and induce the producer to enter the contract with the pipeline, the optimal take percentage will be nonnegative. As a rule, $s(\alpha)$ will equal y only when the producer can sell his output to another pipeline that comparably values the gas, which is unlikely unless the producer is already connected to several pipelines. At the other extreme where gas has no alternative value to the producer, the optimal take percentage is 100 percent, which would be the case, for example, were all of the gas drained away by nearby wells if not extracted by the producer himself.

C. *The Effects of Regulation*

In general, contract terms perform two functions. First, they permit parties to establish a division of the gains from trade that allows both to cover fixed costs; and second,

the buyer may gradually decrease the quantity of gas taken from the producer yields the same qualitative results.

[13] By no net return, we mean zero revenues. Note that a pipeline's willingness to fulfill the terms of a contract may also depend on the opportunity to purchase alternative low-cost gas supplies. In that case, his return in the event of breach would be $\max\{0, \bar{v}\}$ where $\bar{v}$ is his net revenue from the alternative purchases. This possibility does not affect the optimal breach penalty discussed below.

[14] The efficiency of the expectation damage has been demonstrated elsewhere; see, in particular, J. H. Barton (1972), and Steven Shavell (1980). Richard Pierce (1983) has noted that under the Uniform Commercial Code, a pipeline would be liable to a producer for "the difference between the market price and the contract price of the gas available but not taken" (p. 79), precisely the penalty described above. Note that take provisions only address the problem of buyer breach, and as such are only adopted where the principal source of uncertainty is on the demand side of the transaction, a condition true of both coal and gas production. In practice, parties must base their choice of δ on the conditional expectation of $s(\alpha)$ in the event of breach.

they determine the performance incentives in force during execution of the agreement. When contract terms are freely set, transfers between the parties prevent distributional considerations from interfering with the choice of incentive structures.

Thus far we have seen that take obligations serve the second of these functions, and that buyers and sellers may have an incentive to specify such terms even in the absence of regulation and risk aversion. The presence of price controls, however, may distort those incentives. In the case of natural gas, wellhead price restrictions prohibit certain divisions of gains and create an excess demand for gas production capacity. As a result, producers may engage in nonprice competition for capacity through the choice of contract terms. In that regard, raising take obligations will, *ceteris paribus*, increase the amount the seller receives when the buyer rejects delivery, thereby raising the expected value of the contract to the seller.[15] By setting a take percentage greater than the optimal level, the parties in effect sacrifice some efficiency in performance incentives in order to achieve a higher level of investment. The take percentage that would be chosen in a regulated transaction becomes

$$\gamma = 1 - s(\alpha)/y + D, \tag{2}$$

where $D > 0$ reflects the excess demand induced by the presence of binding price constraints.[16]

[15]Because an increase in γ both raises the amount paid in the event of breach and reduces the number of states in which breach occurs, it is possible that beyond some point an increase in γ could, depending on the distribution of θ, reduce the expected value of the contract to the seller. In equilibrium, however, more restrictive price ceilings lead unequivocally to higher take percentages on the margin (see our 1984 working paper).

[16]The additive form of equation (2) follows directly from the model in our earlier paper where we characterize efficient contracts between a producer and a pipeline. The choice of γ in an efficient contract is shown to be a second-best response to the existence of a regulatory price ceiling.

II. Empirical Results

Here we employ data on natural gas producer-pipeline contracts and associated well attributes to test the relationship characterized by equation (2), which we rewrite for estimation purposes as

$$\gamma_i = 1 - s(\alpha_i)/y_i + \xi D_i, \tag{3}$$

where ξ is a coefficient measuring the effects of excess demand, D_i, on take percentages. Prior to estimating this relationship, we describe the data used in the estimation and construct a proxy for D_i.

A. *Gas Contracts and Well Characteristics*

The data used for this study were obtained from several sources. Information regarding price and take obligations was obtained from a survey (EIA-758) conducted by the Energy Information Administration (EIA) in 1982. Through this survey, the EIA obtained detailed data on 659 contracts governing the interstate sale of natural gas from 615 wells located in the lower 48 states. The contracts included in the survey were randomly selected from post-1978 wells which qualified for incentive pricing under Sections 102, 103, 107, and 108 of the Natural Gas Policy Act (NGPA, 1978).[17] Since the data were obtained in disaggregated form, we were able to examine the relationship between take obligations and well characteristics on a well-by-well basis.

Information on the characteristics of the wells governed by these contracts was obtained independently from records at the EIA. Data availability reduced the sample to approximately 300 contracts governing onshore gas priced under NGPA sections 102 ("new" natural gas), 103 ("new" onshore production wells), and 107 ("high cost" gas from deep wells, tight sands, Devonian shales or geopressurized brine).

[17]For a more complete explanation of the survey methodology, see EIA, "Natural Gas Producer/Purchaser Contracts..." (1982).

The variables employed in the estimations are:

γ_i = the contractually specified take percentage for a contract covering gas produced from well i;

$\bar{p}_i$ = the applicable price ceiling (if any) for this gas in October 1981;

p_i^0 = the actual contract price for the gas in October 1981;

D_i = a measure of excess demand for gas from well i;

$DEPTH_i$ = the depth in feet of well i;

$BUYERS_i$ = the number of independent pipelines serving the gas field tapped by well i;

$SELLERS_i$ = the number of independent producers operating in the corresponding field;

$HERF_i$ = the Herfindahl numbers equivalent for the concentration of pipelines in the FERC gas region corresponding to well i,

$$=[\Sigma_t(R_{it}/T_i)^2]^{-1};$$

where R_{it} is the dedicated reserves of interstate pipeline t in FERC gas area i, T_i is the total amount of reserves dedicated to interstate pipelines in the gas area, and the summation is taken over all of the interstate pipelines with reserves in the gas area.[18]

B. *Excess Demand*

As a proxy for the effect of a price ceiling on the demand for gas, let

$$D_i = p_i^* - \bar{p}_i \quad \text{if} \quad p_i^0 = \bar{p}_i;$$
$$= 0 \qquad\qquad \text{if} \quad p_i^0 < \bar{p}_i$$

where p_i^* is the price that would have obtained for well i in the absence of price constraints, and $\bar{p}_i$ is the applicable price ceiling. Thus, if the constraint on price is not binding, $D_i = 0$, and γ is affected only by $s(\alpha)$. If the ceiling is binding, however, D measures the difference between the unconstrained and ceiling prices.

Since p_i^* cannot be observed for $p_i^* \geq \bar{p}_i$, we employ maximum likelihood techniques to construct an estimate for this variable. Where the observations of the dependent variable are truncated, as is the case here, Tobit is an appropriate estimation procedure.[19] In particular, suppose that $p_i^* = aW_i + e_i$, where W_i is a vector of well characteristics affecting p_i^*, and e_i is normally distributed with zero mean and variance σ^2. Then the observed price in a contract covering well i would be

$$p_i^0 = p_i^*, \quad \text{if} \quad p_i^* < \bar{p}_i,$$
$$= \bar{p}_i, \quad \text{if} \quad p_i^* \geq \bar{p}_i.$$

The likelihood function of the ith observation is

$$L_i = f(aW_i) \qquad \text{if} \quad p_i^0 < \bar{p}_i,$$
$$= 1 - F(aW_i) \quad \text{if} \quad p_i^0 = \bar{p}_i.$$

Estimates for p_i^* are derived by maximizing the likelihood function,

$$\Lambda_i = \prod_1^N L_i,$$

with respect to a and σ^2.

Results of the estimation of p_i^* are reported in Table 1. The price at which gas is exchanged in an unregulated transaction depends upon the costs of drilling and connecting the well to a pipeline, as well as on the relative bargaining positions of the transactors. In general, one would expect drilling costs, and hence price, to be positively related to the depth of the well. Also, a large

[18] In order to construct a pipeline, the Natural Gas Act (1938) requires that the transmission company obtain a certificate of public convenience and necessity from FERC. One requirement is that the transmission company demonstrate "adequacy of reserves"; i.e., a sufficient supply of gas to keep the pipeline in operation. As a result, gas reserves are often contractually "dedicated" to individual pipelines. Thus one indicator of the proximity and capacity of pipelines in a geographical region is the number of pipelines that have dedicated reserves in the area and the relative size of those reserves. The information used to compute this measure was obtained from EIA, "Gas Supplies of Interstate Natural Gas Pipeline Companies–1980," (1980).

[19] See, for example, G.S. Maddala (1983).

TABLE 1—ESTIMATION OF p^*

Variable	Coefficient	T-Ratio	Mean
CONSTANT	1.0623	10.890	1.000
DEPTH	$.27465\times10^{-3}$	4.4689[a]	10368.
HERF	.35661	2.7472[a]	4.8687
BUYERS	.34237	2.6665[a]	4.5709
SELLERS	−.06202	−2.4156[b]	9.3041
Chi-square: 403.333 with 4 degrees of freedom[c]			
Number of observations: 296			
Proportion of observations for which $p^0 < \bar{p}$: .1757			

[a] Indicates significance beyond the .01 level.
[b] Indicates significance beyond the .05 level.
[c] Indicates significance beyond the .001 level.

number of buyers and sellers in a particular gas field would tend to undermine the respective group's bargaining position, with corresponding effects on price. In addition, to the extent that connection costs are reduced by the proximity of transmission lines, increases in the concentration of pipelines serving an area should also raise the value of the transaction. The availability of alternative pipelines depends both on their number and transmission capabilities. The Herfindahl measure defined above provides a proxy for the proximity of pipeline capacity to a particular well.

As can be seen from Table 1, the coefficient on each of the variables has the expected sign and is significant beyond the .05 level.

C. *Take-or-Pay Percentages*

We may now turn to estimating the effects of alternative values and excess demand on take percentages. To derive an estimate of $s(\alpha)$, the alternative value of a well, recall that this value will be the maximum of two alternatives: the resource value, which is the discounted value of the gas if left in the ground to be sold at a later date less losses due to drainage by other wells; and the sale value, which is the net value of the gas from sale to another pipeline.

Letting X_i be a vector of attributes affecting the resource value and Z_i a vector affecting the sale value, the expected value of the gas in an alternative use may be represented as

$$s(\alpha_i) = (\beta_1 X_i + \beta_2 Z_i) q_i,$$

where q_i is the capacity of well i, and β_1 and β_2 are coefficient vectors.[20]

Substituting into equation (3) and adding an error term, ε_i, yields

$$(3') \quad \gamma_i = 1 - \beta_1 \frac{X_i}{p_i^0} - \beta_2 \frac{Z_i}{p_i^0} + \xi D_i + \varepsilon_i.$$

The theory of the preceding section predicts that well characteristics, X and Z, which raise (lower) the alternative value of developed capacity should lead to a decrease (increase) in the size of take obligations, and thus implies negative (positive) values of the corresponding coefficients in β_1 and β_2. Meanwhile, we would expect that the more constraining the price ceiling, that is, the greater $p_i^* - \bar{p}_i$, the larger should be γ_i, implying $\xi > 0$.

In determining the resource value of the gas, the larger the number of sellers in a field, the greater the drainage that would occur if a producer were forced to "shut in" his supplies, and hence, the lower this alternative value. The alternative sale value in turn would be expected to increase with the availability of alternative purchasers as measured by the number of buyers in the field and the number of pipelines in the region. Hence, *BUYERS* and *HERF* should both raise the expected alternative value of a well, implying a lower take percentage, and *SELLERS* should reduce $s(\alpha)$ and raise γ.

The ordinary least square (*OLS*) estimates of (3′) are presented in Table 2. (The prefix *S* on a variable indicates that it has been divided by price, see equation (3′). Also note that γ is expressed in percentage terms rather than as a fraction.) Each of the coefficients in this regression has the predicted sign and is

[20] It would be possible to estimate $s(\alpha)$ as the maximum of its resource and sale values using switching regression techniques (see Maddala). We have chosen instead to let well characteristics affect the *expected* value of $s(\alpha)$ additively for its relative computational ease.

TABLE 2—ESTIMATION OF TAKE OBLIGATIONS

Variable	Coefficient	*T*-Ratio	Mean
CONSTANT	82.0822	43.6650	1.000
SSELLERS	.4833	2.0927[a]	3.9819
SBUYERS	−1.3136	−1.7589[b]	2.1080
SHERF	−.8646	−1.7929[b]	2.3270
REGCONST (*D*)	1.85132	3.9956[c]	2.8897

F: 13.24079 with 4 and 294 degrees of freedom[c]
Number of observations: 299
$R^2 = .1526$

[a] Indicates significance beyond the .05 level.
[b] Indicates significance beyond the .1 level.
[c] Indicates significance beyond the .001 level.

statistically significant beyond the 10 percent level. The evidence is thus consistent with the hypothesis that take percentages are designed to influence adaptation in long-term contracts and, in particular, are negatively related to the alternative value of gas reserves. Moreover, price ceilings seem to have the predicted effect on take provisions, raising the percentage paid for nondelivery. The findings indicate that, at the mean, a 1 percent decrease in the price ceiling will lead to a 6 percent increase in the take obligation. For the mean values of the independent variables, the regulatory price ceiling increases the predicted take obligations from 79 to 85 percent.

III. Conclusions

The results of this paper suggest that the incentive to provide flexibility in long-term contracts is an important consideration in the design of contract terms, and that the nature of those terms can be predicted on the basis of characteristics of the transaction. In particular, we have identified an efficiency motivation for the inclusion of take-or-pay provisions in long-term agreements: take obligations induce purchasers to release output to alternative uses only when it is efficient to do so. These incentives may be distorted, however, by the existence of regulated price ceilings. In the case of natural gas, government regulation of wellhead prices appears to have caused nonprice competition in take obligations leading to higher percentages than would prevail in the absence of such regulation.

Empirical tests presented in the paper support the hypotheses of the model. Take percentages are significantly lower for wells associated with small numbers of sellers and large numbers of buyers, each of which raise the alternative value of the gas, a result which seems to apply more generally: E. M. Carney, for example, has noted that, in coal contracts, "If... the coal mine has no access to other markets, the seller obviously has more need for a take or pay clause than he would otherwise"; while, on the other hand, "If the seller can get his product to other markets, the take or pay provision is often tempered to reflect that fact" (p. 226).

The evidence presented here refutes the common perception that take-or-pay provisions are soley an artifact of wellhead price regulation and hence serve no useful purpose in the absence of regulation. Providing that externalities do not result in a divergence of private and social valuations,[21] take obligations contained in contracts written in unregulated environments provide for efficient adaptation to changing circumstances in long-term contractual relationships. Whether a policy to reduce excessive take provisions to *ex post* optimal levels should be advocated on the basis of regulatory interference depends upon both the current regulatory status of the gas covered by the contract and the regulatory environment expected to govern the sale of gas discovered in the future. For example, if the gas under consideration has

[21] Where a field is exploited by a single producer, the gas production process involves no obvious externalities. With more than one operator, however, gas not sold by one producer may be captured and profitably marketed by other producers in the same field. In that event, the social value of "shut in" gas is apt to exceed the value of that gas to the original producer, leading to the adoption of take provisions in excess of socially optimal levels. To the extent that the drainage problem is exacerbated by the number of sellers in the field, the results in Table 2 suggest that take percentages are raised by less than one-half of 1 percent above the socially optimal level for each additional producer in a given field. This implies that, on average, take percentages exceed their socially optimal levels by less than two percentage points due to the externalities associated with competitive extraction from a common pool.

been deregulated and deregulation is expected to continue in the future, the possibility of private bilateral renegotiation obviates the need for intervention. With the removal of the price ceiling, the efficiency gains arising from the reduction of take obligations can be distributed between the buyer and seller through price adjustments, effecting a true Pareto improvement.[22]

Contract terms should also be upheld for categories of gas that currently remain regulated if price regulation is expected to be extended into the future: the precedent established by a reduction of take percentages from the second-best levels stipulated in the contracts could seriously distort capacity investment decisions for future wells. On the other hand, if future discoveries are expected to be sold in an unregulated environment, the efficiency gain from a one-time reduction in take obligations is not offset by adverse precedential effects on future investment. In this event, a prima facie case can be made for intervention to reduce take obligations to levels consistent with efficiency absent regulation. However, these potential benefits must be weighed against the ability of the government to intervene advantageously. Inasmuch as optimal take percentages depend upon characteristics of individual wells, effective intervention would require well by well adjustments. Given the government's previous record on well-specific regulation (see fn. 1), the practicality of this solution is in doubt.

[22] While such renegotiation is certainly not costless, it is likely to be less expensive and more precise than either judicial or legislative intervention.

REFERENCES

Barton, J. H., "The Economic Basis of Damages for Breach of Contract," *Journal of Legal Studies*, June 1972, *1*, 277–304.

Broadman, H. G. and Toman, M. A., "Non-Price Provisions in Long-Term Natural Gas Contracts," Resources for the Future, Washington, 1983.

Canes, M. E. and Norman, D. A. "Analytics of Take-or-Pay Provisions in Natural Gas Contracts," Discussion Paper No. 029, American Petroleum Institute, Washington, 1983.

Carney, E. M., "Pricing Provisions in Coal Contracts," in *Rocky Mountain Mineral Law Institute*, New York: Matthew Bender, 1978, 197–230.

Goldberg, Victor and Erickson, John, "Long Term Contracts for Petroleum Coke," Department of Economics Working Paper No. 206, University of California-Davis, September 1982.

Harris, Milton and Townsend, Robert, "Resource Allocation Under Asymmetric Information," *Econometrica*, January 1981, *49*, 33–64.

Hawkins, C. A., *The Field Price Regulation of Natural Gas*, Tallahassee: Florida State University Press, 1969.

Klein, Benjamin, Crawford, R. G. and Alchian, Armen A., "Vertical Integration, Appropriable Rents and the Competitive Contracting Process," *Journal of Law and Economics*, October 1978, *21*, 297–326.

Libecap, G. D. and Wiggins, S. N., "Contractual Responses to the Common Pool: Prorationing of Crude Oil Production," *American Economic Review*, March 1984, *74*, 87–98.

MacAvoy, Paul W. and Pindyck, Robert S., *The Economics of the Natural Gas Shortage (1960–1980)*, New York: Elsevier, 1975.

McDonald, S. L., *Petroleum Conservation in the United States: An Economic Analysis*, Baltimore: Johns Hopkins University Press, 1971.

Maddala, G. S., *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press, 1983.

Masten, S. E. and Crocker, K. J., "Regulation and Nonprice Competition in Long Term Contracts for Natural Gas," Working Paper, University of Virginia, December 1984.

Norman, D., "Indefinite Pricing Provisions in Natural Gas Contracts," Discussion Paper No. 034, American Petroleum Institute, Washington, 1984.

Pierce, Richard J., "Natural Gas Regulation, Deregulation, and Contracts," *Virginia Law Review*, 1983, *68*, 63–115.

Shavell, Steven, "Damage Measures for Breach of Contract," *Bell Journal of Economics*, Autumn 1980, *11*, 466–90.

Sullivan, R. E., *Handbook of Oil and Gas Law*, New York: Prentice-Hall, 1955.

Williamson, Oliver E., "Transaction-Cost Economics: The Governance of Contractual Relations," *Journal of Law and Economics*, October 1979, *22*, 233–62.

Energy Information Administration (EIA), "Natural Gas Producer/Purchaser Contracts and Their Potential Impacts on the Natural Gas Market," DOE/EIA-0330, Washington: USGPO, 1982.

______, "Gas Supplies of Interstate Natural Gas Pipeline Companies—1980," DOE/EIA-0167 (80), Washington: USGPO, 1982.

Mines and Migrants in South Africa

By Robert E. B. Lucas*

Between 1971 and 1978, wages of more than one-half million nonwhite laborers in the South African mines tripled in real terms. In the same period, the nonwhites employed in the mines switched from being 62 percent foreign to 62 percent domestic.[1] These changes followed a period—from 1911 to 1971—during which real wages of black gold miners did not rise, and terminated almost a century of reliance on foreign labor "reserves" for the majority of such labor.[2]

These dramatic events are examined here in the context of an econometric model of the demand for labor by the South African mining sector from 1946 to 1980. This affords an unusual opportunity to study the demand side of a market for internal and international migrants, in a society where racial discrimination is formalized in the apartheid system, where powerful mining houses wield potential monopsony power, and where political factors in the region are major determinants of economic behavior.

To comprehend the derived demand for workers in this sector, it is essential to outline at least certain aspects of the industry's organization and that of the market for labor; this is undertaken in Section I. Section II develops a stylized model, which is then estimated, from data described in Section III, for the gold, diamond, coal, and other minerals sectors separately in Section IV.

*Associate Professor of Economics, Boston University, Boston, MA 02215. Too many people have contributed to this work through interviews and discussions to mention and thank each individually—to all I am most grateful. The initial phase of this study was supported by the International Labour Organization, but opinions and interpretations herein expressed are mine alone.

[1] "Nonewhite" comprises those designated black, colored, and Asian. The term "domestic" (or South African) includes the several black states or homelands, such as Transkei and Bophuthatswana.

[2] See Francis Wilson (1972), W. J. Breytenbach (1979), and Merle Lipton (1980).

I. Organization of the Mine Labor Market

Ninety percent of the 700,000 workers employed in the South African mines in 1980 were nonwhite. Most were employed in mines affiliated with the Chamber of Mines, one of whose chief functions is to organize hiring and wage policy for the sector. For example, the Chamber—an organization of employers without government or labor representation—for many years imposed on members a "maximum permissible average" wage they might offer nonwhites and today effectively imposes a minimum also. Almost all gold, diamond, and coal production (employing 431, 19, and 80 thousand nonwhites, respectively, in 1980) is by mines affiliated with the Chamber. Some extraction of other minerals (employing 109 thousand nonwhites) is also by affiliated mines and several of the nonaffiliated mines are owned by the half-dozen mining houses that dominate the Chamber. Nonwhite labor is recruited for members of the Chamber by licensed agencies; the largest of which is now called The Employment Bureau of Africa (TEBA). Although the diamond sector is affiliated with the Chamber, it is not a member for labor purposes and this has had profound effects for wages in this sector to be explored later. Moreover, only one mine in the other minerals sector employs TEBA to recruit its labor. Naturally, the centralization of hiring in recruiting agencies, subject to the decrees on wages set by consent in the Chamber, places the industry in a potentially monopsonistic position if faced with an upward-sloping supply of labor.

Foreign workers are hired on contracts varying in length from six months to two years, according to country of origin. None is permitted to bring his family with him and at the end of their contract, workers are required to return to their own country. Not surprisingly this system—designed in part to

allow men to return for ploughing and consequently not put pressure on wages by rendering the family entirely dependent on mine earnings, and partly to avoid more permanent settlement of foreign blacks albeit without their families—generated very high turnover rates of labor in the sense of workers not returning from their compulsory home visit. Until the mid-1970's, black workers could not elect to quit a contract (illegal quitters being termed deserters); that is, not until a boycott by American unions of South African coal imports, on the grounds they were produced by indentured labor, prompted the abolition of the "master and servants" acts. Since then, the fixed-period contracts have, of course, dwindled in importance though nominally they still exist.

In 1973, Malawi and Mozambique were the two largest suppliers of foreign labor, providing 39 and 27 percent, respectively. In 1974, President Hastings Banda cut off further recruiting of mine workers from Malawi —ostensibly because of a plane crash that was full of recruits. Since miners from Malawi were recruited on two-year contracts, the effect took two years to work through, but the number of Malawians employed in the mines fell from 140,000 in 1974 to 2,000 in 1977. At almost the same instant, the FRELIMO government finally succeeded in ousting the Portuguese. Whether the subsequent sharp decline in Mozambican recruits reflected the demands of the South Africans or exigencies of the new regime remains a matter of debate (see Ruth First et al., 1977). But overall there existed, after 1974, an excess supply of foreign workers as mine wages rapidly doubled and tripled. In Botswana and Lesotho, many experienced men are turned away. Very few foreign novices are now taken and even current, experienced workers are generally only rehired after their compulsory annual home return if issued with a Valid Reengagement Certificate (VRC). The VRCs are issued to about 70 percent of workers, subject to "good" performance, but obviously the rate of issuance may be varied to control numbers. Thus, when Malawi again lifted the ban on recruiting in 1978, an excess supply of Malawian recruits emerged for a now limited number of openings.

South African black workers are broadly bifurcated into those with and without section 10 rights.[3] Those without must return to their designated homeland each year and must leave their family behind while working in a white prescribed area. Section 10 workers are on average better trained, more urbanized, and better educated in a society where educational policy towards blacks has handicapped industry through lack of skilled labor. In 1970, average earnings of nonwhites in manufacturing were exactly three times those in gold mining. Combined with atrocious working conditions, the social stigma attached to mining in South Africa, and the separation from family and potential loss of section 10 rights ensuant upon relocation, this meant mining attracted virtually no section 10 workers who naturally preferred manufacturing. Employers in manufacturing also preferred section 10 workers because of their skills and because they were required to hire labor within their own Labour District rather than importing domestic or foreign migrant workers.

The labor force without section 10 rights faces four alternatives: to gamble on being allocated a manufacturing job through the Labour Bureaux; to squat as an agricultural worker at a rate of pay which for regular employees even in 1969 was well below the mining wage; to practice subsistence farming on the exceedingly poor lands alloted to nonwhites; or to be recruited by TEBA for mining. As mine wages rose in the 1970's, the last alternative became relatively more attractive, although wages in agriculture and manufacturing, and employment in the latter also rose somewhat. For example, the Zulus, whom recruiters believed would never work

[3]"Section 10" refers to section 10 of the Bantu Urban Areas Act of 1945, which prohibited blacks from remaining in the "white" areas for more than 72 hours unless they had resided there continuously since birth, had worked continuously for the same employer for 10 years, or for more than one employer for 15 years. See Lipton, who reports that 42 percent of black workers have section 10 rights.

below ground, now do so. Indeed, by 1979, even South African labor was in excess supply to mining. In large part, this is attributable to forced relocation of massive numbers of squatters from white farms to homelands in the late 1970's, nominally for security reasons, expanding the pool available to mining and rapidly diminishing the marginal product in subsistence farming.

Unlike nonwhites, white mine workers have been unionized throughout this period, and average earnings of whites in gold mining in 1970 were over 20 times greater than those of nonwhites. Under the apartheid system a rigid color bar is imposed on access to certain occupations, a regime vehemently defended by the (white) Mine Workers' Union. Even so there has been some realignment of job tasks within occupations, a part of which was formalized in an agreement between the mines and the Mine Workers in 1973, but which has progressed informally since.

The 1970's also witnessed unprecedented increases in mineral prices. After 1971 the producer price of gold began to rise relative to the South African cost of living index. By 1974, the real price was 2.9 times the 1971 level and by 1980 the increase was 6.7 fold, though the path was not monotonic in the interim. Over the same period, real diamond prices increased 2.7 times, though 1980 prices remained below the 1946–51 average in real terms, 1971 having been the low point in a steady decline. Coal prices are regulated in three categories: the price paid by the electricity authority is negotiated bilaterally; sales to the remaining domestic market are at administered prices; and exports fetch a much higher price but are effectively limited by port facilities. The average producer price of coal rose only gradually in real terms from 1946 to 1973, then increased 2.4 fold by 1979 as the administered price was raised. Prices of other minerals in 1980 were on average about equal to those in 1969 in real terms, though a very slight rise occurred in between, and remained well below the real prices prevailing in the Korean War era.

Finally, before turning to the model, some of these elements may be pulled together to emphasize the sequence of events. Apart from diamonds, the rest of the mining sector behaved fairly uniformly from 1946 to 1971. After 1971, the Chamber decided to begin raising wages of nonwhites, linking this with stabilization of the labor force in the sense of reduced turnover.[4] The motivations for raising wages initially were at least threefold: a concern for the publicity in both the domestic and world press being given to the very low wages then paid; an efficiency wage argument, enhancing productivity through nutrition and health, the prevailing rates often being described as starvation wages; and the hope of reducing turnover and attracting more skilled workers. In an inherently dangerous industry, and one where security from theft (especially in diamonds) is vital, a stabilized labor force is at a premium, though possibly at the cost of increasing organization and politicization. The initiative to raise wages apparently stemmed from the Anglo American Corporation and particularly from new directors appointed about 1971, but was rendered possible only in an environment of rising gold prices after 1971 and of coal subsequently.[5] In the diamond mines, these changes began much earlier—from 1964 onwards—despite declining real diamond prices at the time. In that instance, the personal role played by Harry Oppenheimer cannot be discounted, nor the role of security risks. From that time onwards, wage policy of the diamond sector was independent of that of the Chamber, the diamond mines not being affiliated for labor purposes, and wages were well above those in the rest of the industry.

After 1974, an era of excess supply of foreign labor follows. Though not everyone in the industry agrees, this probably resulted from a desire to reduce dependence on foreign labor.[6] Such a desire could have been

[4] Lipton uses the term stabilize in this context to mean to settle permanently with one's family. The industry's interest in this form of stabilization is far more ambivalent.

[5] The operations in which Anglo American was involved in the Zambian copper belt had their labor force localized and stabilized some 10 years before the turning point in South Africa (see Philip Daniel, 1979).

[6] In the first 11 months of 1981, 15.88 percent of all TEBA recruits were novices, but for no foreign country

induced by the uncertainties stimulated by the Malawi cut off in 1974, the FRELIMO takeover in 1975, and or the mine compound disturbances of 1975 (reputedly leaving some 200 dead) prompted by a compulsory deferred pay scheme introduced by Lesotho. To attract sufficient South African workers, wages continued to rise.

After 1979, the mines were at full complement of labor and South African workers in excess supply to the industry as chronic underemployment in the homelands was well established under compulsory relocation. From 1978 through 1980, real wages of nonwhites remained almost constant. The industry is now split. Some mining houses wish to raise wages further; others want to lower them again, with the realization the latter would mean more employment. The South African government has encouraged the wage increases in the industry. Though concerned with the economic viability of the homeland situation they have created, their enthusiasm for expanded employment of blacks in mining is not clear, for with the exception of Bophuthatswana, the mines are too far from the homelands. The government would prefer to see employment created along the peripheries of the homelands, for the primary concern is protection of the exclusive white environs.

II. A Stylized Model

A. *General Form*

The story may be broken into two major periods—before and after 1974. In the earlier era, potential monopsony and job discrimination are major features. In more recent times, excess supply of foreign labor and a possible preference for reduced dependence on foreign labor are added.

did this rate exceed 10 percent, and for Mozambique and Malawi it was less than 1 percent. Given the obvious, excess supply of foreign novices, this strongly supports the notion of bias against foreign dependence. Indeed, the stated industry policy is to take no Mozambican novices for fear of political activism in the mine compounds.

1. *Prior to 1974.* Before 1974 it is assumed no excess supply of nonwhite labor to the mines existed either within South Africa or from abroad. To allow for the possibility that the combined supply ($f+s$, where f is foreign labor supply, s is South African supply) is less than infinitely elastic with respect to the wage for nonwhites (w), the supply function on which the sector operates is written

$$f+s=F(w). \tag{1}$$

Job segregation under apartheid, substantially different access to education and training for nonwhites, enforced migrancy resulting in high turnover, and low wages generating poor health and nutrition, all act to keep productivity of nonwhites below that of whites. Recognizing this distinction between nonwhite (b) and white (l) labor employed, output (q) is expressed as some function:

$$q=q(b,l,k), \tag{2}$$

where k is capital stock.

With the exception of diamonds, it is assumed that producers are price takers in the commodity market. The flow of gold is small compared to the world stock, and, as a first approximation, it is not unreasonable to imagine prices determined by international portfolio behavior rather than through flow supply and demand. The producer price of coal has also been regulated throughout the period under consideration as outlined in Section I.

In acting as a potential monopsonist but not monopolist, it is assumed the nondiamond sectors, at least, have been willing to hire nonwhite labor up to the point where marginal cost equals value of marginal product subject to the production function (2). In other words, discrimination against nonwhites is confined to job discrimination as reflected in the production function rather than to additional wage discrimination.[7] In

[7]In addition, complications arising from exhaustible resources are suppressed for simplicity, which is rela-

fact, this is probably a very plausible assumption, for if anything the concern of the mining houses has been to reduce even job discrimination as reflected in their 1973 agreement with the white Mine Workers' Union noted earlier, and in the revisions in job tasks. Nonetheless, some results on tests for an extended model incorporating wage discrimination are cited in Section IV.

The wages (v) of white workers are determined through bargains struck with the Mine Workers' Union and hence are taken as exogenous to this model. The first-order conditions with respect to both types of labor therefore yield

$$p \cdot q_b = w'(f+s) \cdot b + w, \tag{3}$$

$$p \cdot q_l = v, \tag{4}$$

where p = price of output, q_j = marginal product of factor j, and $w'(f+s)$ = partial derivative of w with respect to $f+s$.

The model for this first stage is then closed by assuming

$$f+s=b. \tag{5}$$

2. *After 1974.* In the second stage, three major changes in the model are necessary: a preference not to rely on foreign labor is allowed for, as a result foreign labor is in excess supply, but this is accompanied by a more stable labor force that may be more productive. To characterize this phase, a supply function of domestic labor to the mining sector must be distinguished.

$$s=s(w). \tag{6}$$

Moreover, the production function requires modification for reasons to be discussed in Part B below, but for now we may simply write this as

$$q=Q(b,l,k). \tag{2'}$$

First-order conditions must now be distinguished for domestic and foreign labor separately, thus for domestic, foreign and white labor, one obtains

$$p \cdot Q_s = w'(s) \cdot b + w \tag{3'a}$$

$$p \cdot Q_f = w + \Psi \tag{3'b}$$

$$p \cdot Q_l = v, \tag{4'}$$

where Q_j = marginal product of factor j, $w'(s)$ = partial derivative of w with respect to s, and Ψ = a coefficient of preference not to rely on foreign labor.[8]

It is intriguing to note that this preference against foreign labor has not taken the form of wage discrimination. Foreign and domestic labor continue to be paid equal wages. A potential reason might be tension in the compounds between miners living and working together but rewarded differentially. Yet this friction might be avoided by means of a foreign remittance or recruitment tax. For such taxes there are precedents, though revenue, collected by the mines or recruiting agencies, has gone to the sending countries. I believe the true reason for maintaining equal wages has stemmed from the initial cause of pay increases—sensitivity to accusations of paying exploitative wages.

B. *Specific Form*

To estimate this system, specific forms of the production functions and nonwhite labor supply functions must be imposed. The overall labor supply function (1) is assumed to have constant wage elasticity η and is specified as

$$\ln b = \gamma_0 + \eta \cdot \ln w + \Gamma \cdot X + \gamma_1 \cdot \ln b_{-1} \tag{7}$$

where b_{-1} is b lagged by one year, a term

tively justifiable in view of the massive reserves of most minerals.

[8]Since foreign labor is in excess supply and hence infinitely elastic at the going wage, an objective function for the sector consistent with (3′a), omitting white labor and capital for brevity, would be

$$\Omega = p \cdot Q(f,s) - w(s) \cdot (f+s) - \Psi \cdot f.$$

In other words, the mining houses act as if the cost of hiring foreign labor exceeds the wage rate by an amount Ψ per foreigner.

included to reflect the effects of the two-year contracts prevalent in some of the major supplying countries, remembering that such contracts were binding prior to repeal of the masters and servants acts. The X is a vector of principal components of exogenous elements, both abroad and within South Africa, likely to have shifted the supply function.[9]

As noted in Section I, jobs in the South African manufacturing sector have paid nonwhites much better wages than mining. But availability of the former has been limited, particularly for those without section 10 rights who are the really potential mine workers. The supply of South African nonwhite labor to the mining sector is therefore made a function of the expected wage in manufacturing (m). In addition, supply is likely to rise with the size of the nonwhite male labor force (n). Given this, a constant wage elasticity (σ) specification of (6) is again adopted:

$$(8)\quad \ln s = \sigma_0 + \sigma\cdot \ln w + \sigma_1\cdot m + \sigma_2\cdot n + \sigma_3\cdot \ln s_{-1}.$$

Capital inputs are in fact divided into two types—fixed capital (k) and equipment (e). In selecting a form of the production function for derived demand analysis, it is important to allow for substitution between factors. Four factors of production (b, l, k, and e) and their interactions necessarily imply a lengthy formulation, and within these general requirements two specifications are explored—a translog and a quadratic. (For a discussion of the properties of each see Melvin Fuss, Daniel McFadden, and Yair Mundlak, 1978.) The quadratic form is

$$(9)\quad q = \alpha + \beta\cdot b + \lambda\cdot l + \kappa\cdot k + \varepsilon\cdot e + \beta\lambda\cdot b\cdot l + \beta\kappa\cdot b\cdot k + \beta\varepsilon\cdot b\cdot e + \lambda\kappa\cdot l\cdot k + \lambda\varepsilon\cdot l\cdot e + \kappa\varepsilon\cdot k\cdot e + 1/2(\beta\beta\cdot b^2 + \lambda\lambda\cdot l^2 + \kappa\kappa\cdot k^2 + \varepsilon\varepsilon\cdot e^2) + \beta\tau\cdot s\cdot T + \tau\cdot T.$$

The last two terms require some additional explanation. T is a dummy variable turned on after 1974. Thus $\beta\tau$ reflects any change in productivity of nonwhite labor in the later period, as the labor force is localized with the intent of establishing a more stable, experienced work force. But this period also experienced very sharp rises in mineral prices, and some mines began to rework previously abandoned deposits and several open-cast coal mines commenced operation, each with inherently different productivity—hence the term τT. The translog version of (10) is of course identical except q, b, l, k, e, and s are each replaced by their natural logarithms.

Solving the first-order conditions subject to these specific forms provides for the quadratic case:

$$(10)\quad b = \frac{-\beta}{\beta\beta} + \frac{(1+\eta)}{\eta\cdot\beta\beta}\cdot\frac{w}{p}\cdot(1-T) + \frac{1}{\beta\beta}\cdot\frac{w}{p}\cdot T + \phi\cdot T - \frac{\beta\lambda}{\beta\beta}\cdot l - \frac{\beta\kappa}{\beta\beta}\cdot k - \frac{\beta\varepsilon}{\beta\beta}\cdot e$$

$$l = \frac{-\lambda}{\lambda\lambda} + \frac{1}{\lambda\lambda}\cdot\frac{v}{p} - \frac{\beta\lambda}{\lambda\lambda}\cdot b - \frac{\lambda\kappa}{\lambda\lambda}\cdot k - \frac{\lambda\varepsilon}{\lambda\lambda}\cdot e$$

where ϕ expresses the preference not to rely on foreign labor in units of labor and is assumed constant, so that in the quadratic case $\Psi = p\cdot\beta\beta\cdot\phi$. A negative value of ϕ indicates a preference against foreign labor; a zero value implies no such preference. Comparing (3′a) and (3′b), it is readily seen that $\psi = w'\cdot b + p\cdot(Q_f - Q_s)$, so that if no monopsony power were exercised vis-à-vis South African labor in this period, and if

[9]The 23 underlying variables are: population and rainfall in Botswana, Lesotho, Malawi, Mozambique (rainfall excluded), South Africa, and Swaziland; the expected domestic wage in Botswana and Lesotho, the price of maize, cotton, and tobacco in Malawi; plantation employment, industrial employment and wage, and marketed crops in Mozambique; employment and wage for blacks in South African manufacturing; and a dummy for independence of Malawi. The demand model presented in this paper is part of my larger study (1983) incorporating also the supply side of the story, in which these 23 variables receive explicit treatment.

foreign and domestic workers were perfect substitutes, ϕ would be zero.

In the context of the translog production functions, all terms in b, l, k, and e are replaced by their natural logarithms, and w and v are multiplied by their respective own-factor amount over output.

The system of equations (10) assumes instantaneous adjustment. A simple, flexible accelerator may, however, be appended by adding a lagged dependent variable term to each equation. The well-known interpretation is that all terms in (10) are thereby multiplied by one minus the flexible accelerator coefficient.[10]

Finally, as mentioned earlier, the model is to be estimated for each of the gold, diamond, coal, and other minerals subsectors separately. For gold and coal, (10) is appropriate, but, for the other two, some modification is necessary. The extraction of other minerals is almost entirely by mines that do not recruit through TEBA and commonly not members of the Chamber. It is unlikely these mines have restrained their hiring to help the monopsony position of the industry as a whole, so they are assumed to have behaved as though η were infinite.[11]

Section I notes that the diamond sector is not a member of the Chamber for labor purposes and has pursued an independent wage policy with stabilization of the nonwhite labor force commencing ten years earlier. To explore the longer-term productivity consequences of this stabilization and its implication for the role of white labor, the last two terms of the production function (2) are replaced by $T\cdot(\beta\tau\cdot b\cdot t\cdot+\lambda\tau\cdot l\cdot t+\beta\lambda\tau\cdot b\cdot l)$, where T is now a dummy for the post-1964 period and t measures time. Thus, $\lambda\tau$ and $\beta\tau$ reflect progress in productivity of white and nonwhite labor with the passage of time and gaining of experience under stabilization; $\beta\lambda\tau$ reflects any shift in labor substitutability after diminution in job barriers. In addition, the possibility of exploitation of monopoly power by DeBeers warrants incorporation. Thus, if δ represents the elasticity of demand for diamonds, rearrangement of the first-order conditions for nonwhite and white labor in diamonds provides, in the quadratic case:

$$(11)\quad b=\frac{-\beta}{\beta\beta}+\frac{1}{(1+1/\partial)\cdot\beta\beta}\cdot\frac{w}{p}-\frac{\beta\lambda}{\beta\beta}\cdot l$$

$$-\frac{\beta\kappa}{\beta\beta}\cdot k-\frac{\beta\varepsilon}{\beta\beta}\cdot e+T\cdot\left[-\frac{\beta\tau}{\beta\beta}\cdot t-\frac{\beta\lambda\tau}{\beta\beta}\cdot l\right]$$

$$l=\frac{-\lambda}{\lambda\lambda}+\frac{1}{(1+1/\partial)\cdot\lambda\lambda}\cdot\frac{v}{p}-\frac{\beta\lambda}{\lambda\lambda}\cdot b$$

$$-\frac{\lambda\kappa}{\lambda\lambda}\cdot k-\frac{\lambda\varepsilon}{\lambda\lambda}\cdot e+T\cdot\left[\frac{-\lambda\tau}{\lambda\lambda}\cdot t-\frac{\beta\lambda\tau}{\lambda\lambda}\cdot b\right]$$

Estimates of equations (9), (10), and (11) are discussed in Section IV.

III. The Data

The model is estimated on annual data from 1946 to 1980. Space does not permit a full discussion of the sources and compilation of data assembled for this study; however, a few words are in order.

Output by each of the subsectors is measured in physical terms (ounces of gold; 1000 carats of diamonds; tons of coal). For the other minerals, a quantity index is adopted (actually several published Divisia indices spliced together) with base in 1957 at 1000. Employment is simply the average numbers at work during the year. The two types of capital stock (k; e) are measured in constant, 1970 1000 rands. The stock figures are derived from gross expenditures deflated by the South African cost-of-living index, with k depreciated linearly at 4 percent and e at

[10]See M. Ishaq Nadiri and Sherwin Rosen (1973). Extension to a nondiagonal matrix of coefficients therein proposed or to decision making when lags are constraints, as in Hiroshi Yoshikawa (1980), is left for later work.

[11]In the post-1974 period, the other minerals sector has continued to follow the wage leaders partly for the same reasons that motivated the initial raises by the Chamber (particularly in those mines within the other minerals sector owned by the large mining houses), and partly because many of these mines lacked the organization to recruit foreign labor and local labor required higher wages.

10 percent. Capital stock k incorporates mining property, shafts, and mine development, and e is simply equipment. Wages are measured by earnings per employee per year. Output prices are derived from realized value of sales divided by sales—rands per ounce of gold, per carat of diamonds, per ton of coal, and 1000 rands per index unit of output for other minerals. The South African nonwhite male labor force is proxied by the nonwhite male population ages 18 to 49, measured in millions of men. The expected wage in manufacturing is then nonwhite employment in manufacturing, construction, and South African Railways and Harbours, relative to the male labor force, multiplied by annual average earnings of blacks in manufacturing deflated by the cost of food index.

IV. Estimation of the Model

All equations are estimated by three-stage least squares using Fair's technique with respect to instrumental variables, unless otherwise noted. The theory suggests that a large number of nonlinear cross-equation constraints might be imposed on the estimates. However, no such restrictions are imposed here except in the context of hypothesis testing.

A. *Monopsony*

From the equation for nonwhite labor in (10) it is readily seen that the coefficients on wage before and after 1974 differ only by inclusion of $(1+\eta)/\eta$ in the former, where η is the elasticity of labor supply. If the sector tends towards behavior consistent with an infinitely elastic supply, rather than as a monopsonist, $(1+\eta)/\eta$ should approach one and the coefficients on the two-period wage terms become equal. As discussed in Section II, this hypothesis is to be tested with respect to the gold and coal sectors, neither the diamond nor other minerals sector having been part of the Chamber for labor purposes. The estimates of nonwhite labor input equations are presented in Table 1, with t-statistics for a zero null-hypothesis presented in parentheses beneath each coefficient.

In each context—gold and coal, quadratic, and translog—the wage terms have negative estimated coefficients with a larger absolute value in the earlier period. Moreover, in each case except that of coal in translog form, this difference between time periods is statistically significant at more than a 90 percent confidence level.

But to argue that these results reflect exploitation of monopsony power, at least in gold, one must proceed further. Was the supply of labor facing the industry indeed upward sloping in the earlier period before 1974? Table 2 presents an estimate of the overall labor supply equation (7) up to 1974.[12] Although the region as a whole is frequently referred to by the industry as a labor reserve, the estimate of (7) shows an upward-sloping supply of labor with elasticity even significantly less than one at a 99 percent confidence level. Merle Lipton notes:

> It had long been argued by the industry that Black peasants have a backward-sloping supply curve...they were target workers.... The fact that this argument was a convenient rationalization for low wages was nowhere more clearly demonstrated than by the action of the Chamber, in the wake of the gold price rise, in raising wages to increase supply. [p. 109]

The result in Table 2 is certainly consistent with Lipton's view. Furthermore, the estimate of η reported in Table 2 is not significantly different from that implied by the estimated nonwhite labor input equations for gold, when η is obtained as a nonlinear transformation of the wage coefficients in the two periods (Ernst Berndt et al., 1974).

But this latter derivation of η and, indeed, the prior tests for monopsony are founded on the theory that the coefficients in Table 1 on the wage term after 1974 equal $\beta\beta$. This

[12] This estimate is obtained by two-stage least squares, using Fair's technique. The equation includes 11 principal component terms in the vector X, selected by backward stepwise regression, not explicitly shown in Table 2.

TABLE 1—NONWHITE LABOR INPUT EQUATIONS

	Quadratic				Translog		
	Gold	Coal	Diamonds	Other	Gold	Coal	Diamonds
Intercept	154538	53835	−5113	11319	6.83	9.88	.563
	(2.00)	(9.28)	(1.67)	(0.74)	(2.46)	(10.40)	(0.32)
Wage Term			−118	−14850			−.0017
			(5.94)	(2.70)			(2.37)
Wage Term, Pre-1974	−8774	−139			−2.73	−1.82	
	(2.43)	(2.63)			(1.74)	(1.85)	
Wage Term, Post-1974	−2647	−1.60			−.424	−1.34	
	(1.50)	(0.03)			(1.11)	(1.29)	
Post-1974 Dummy	−74207	−18136		12519	−.217	−0.43	
	(2.41)	(1.75)		(1.68)	(1.71)	(0.23)	
White Labor	2.27	5.32	3.55	11.5	.280	.430	.559
	(1.53)	(11.35)	(3.60)	(9.03)	(1.50)	(5.22)	(2.24)
Equipment	−.029	−.051	.101	0.65	−.008	−.136	.266
	(0.37)	(2.94)	(3.97)	(0.67)	(0.07)	(2.46)	(2.75)
Fixed Capital	.220	−.089	−.113	−.139	.246	−.068	−.062
	(4.56)	(2.59)	(3.56)	(1.09)	(1.56)	(0.87)	(0.40)
Time, Post-1964			669				.110
			(4.19)				(0.21)
White Labor, Post-1964			−2.99				−.024
			(3.29)				(0.13)
Lagged Dependent Variable			.470				.247
			(6.08)				(1.22)
Rho	.385	.424	−.001	.816	.419	.457	.614
Standard Error of *Rho*	.155	.168	.192	6.46	.148	.162	.189
Number of Observations	33	33	33	33	33	33	33

may be tested directly by viewing the production functions, and estimates for the case of gold are shown in Table 3.[13] Therein, both the translog and quadratic estimates provide a negative coefficient on the square of nonwhite labor. Following A. Ronald Gallant and Dale Jorgenson's (1979) test for nonlinear, cross-equation restrictions, this production function estimate of $\beta\beta$ is not significantly different from the implied value estimated in the nonwhite labor input equation.

All of the evidence is thus consistent with exploitation of monopsony power, at least by the dominant gold sector, before 1974. Indeed, this seems eminently sensible, providing a rationale for the centralized system of recruiting and imposition of the maximum permissible average wage payable by mem-

TABLE 2—NONWHITE LABOR SUPPLY EQUATIONS

	Foreign and Domestic Labor 1948–73	South African Labor 1947–78
Intercept	−3.22	1.97
	(2.86)	(1.70)
Log Wage	.140	.498
	(4.12)	(3.69)
Expected Wage in Industry		−46.0
		(2.31)
Labor Force		.0303
		(1.74)
Lagged Dependent Variable	.848	.726
	(29.78)	(6.81)
Durbin's h-Statistic	0.18	0.92
Number of Observations	26	32

Note: Also included in the first equation, but not shown, are 11 principal component terms for shifts in supply.

[13] No correction for first-order serial correlation is made in these estimates. If *rho*, a serial correlation coefficient, is estimated, the associated t-statistic for a zero null hypothesis is 0.57 for the quadratic case and 1.07 for the translog.

TABLE 3—PRODUCTION FUNCTIONS

	Gold		Diamonds	
	Quadratic	Translog	Quadratic	Translog
Intercept	$-.14\times10^9$	-1055	-121962	-677
	(1.16)	(1.62)	(1.77)	(0.58)
Nonwhite labor	-755	-257	4.21	261
	(1.39)	(1.89)	(2.55)	(2.61)
White labor	9417	340	48.6	-4.35
	(1.61)	(2.07)	(1.24)	(0.01)
Capital	331	229	$-.111$	41.1
	(1.41)	(2.10)	(0.35)	(1.94)
Equipment	-158	-101	.313	-146
	(0.48)	(1.61)	(0.38)	(2.31)
Nonwhite × White	.032	34.7	$-.001$	11.3
	(2.78)	(3.32)	(0.90)	(0.86)
Nonwhite × Capital	.00084	20.8	$-.25\times10^{-4}$	-1.21
	(2.08)	(2.84)	(1.86)	(0.63)
Nonwhite × Equipment	$-.0012$	-15.5	$-.54\times10^{-4}$	-3.38
	(1.91)	(2.59)	(2.22)	(0.62)
White × Capital	$-.010$	-31.2	$.18\times10^{-3}$	2.18
	(1.64)	(2.45)	(1.53)	(0.59)
White × Equipment	.010	19.1	$-.13\times10^{-4}$	.474
	(1.17)	(2.20)	(0.07)	(0.06)
Capital × Equipment	.00048	14.6	$-.20\times10^{-5}$	-4.22
	(1.49)	(2.56)	(0.43)	(3.02)
Nonwhite^2	$-.0011$	-7.63	$-.29\times10^{-4}$	-16.2
	(1.48)	(1.38)	(0.46)	(3.50)
White^2	$-.204$	-28.4	$-.005$	-7.65
	(2.51)	(2.98)	(0.96)	(0.35)
Capital^2	$-.00023$	-12.8	$-.15\times10^{-5}$	$-.145$
	(1.94)	(2.86)	(1.41)	(0.39)
Equipment^2	$-.00023$	-3.98	$.11\times10^{-5}$	10.4
	(0.88)	(2.22)	(1.40)	(4.06)
Domestic Labor, Post-1974	-24.1	$-.242$		
	(1.82)	(2.05)		
Post-1974 Dummy	$-.17\times10^7$	2.74		
	(0.55)	(1.88)		
Nonwhite × Time, Post-1964			.092	2.64
			(8.20)	(5.10)
White × Time, Post-1964			$-.265$	-2.96
			(5.01)	(4.74)
Nonwhite × White, Post-1964			$-.14\times10^{-3}$	$-.042$
			(2.56)	(0.82)
Durbin-Watson Statistic	2.13	2.16		
Rho			$-.311$	$-.669$
Standard Error of *Rho*			.231	.273
Number of Observations	33	33	33	33

ber mines. as outlined in Section I. In addition, the recruiting system appears to have provided the mining industry with an ability to respond quickly to changing situations. Thus, if lagged dependent variables are inserted in the nonwhite labor equations for gold and coal, the flexible accelerators prove statistically indistinguishable from zero. These terms are consequently omitted from the equations reported in Table 1. In contrast, the combination of up to two-year contracts plus the master and servants laws, prohibiting legal quitting, have rendered the ability of workers to adapt to changing conditions very sluggish. The coefficient on the lagged dependent variable

in the overall labor supply equation in Table 2 is thus not only significantly positive, but indeed very large.

B. *Preference for Domestic Labor*

In (10) a negative coefficient ϕ on the post-1974 dummy represents a preference not to rely on foreign workers. In Table 1, this coefficient is estimated to be significantly less than zero at a 95 percent confidence level in both formulations for gold, and for coal in the quadratic case. Indeed, the effect is quite large. Thus, the quadratic estimates suggest that, despite the escalated wages of this period, the gold sector would have wanted to employ some 70,000 more nonwhites and the coal sector nearly 20,000 extra as mineral prices rose, had a policy of switching to domestic labor not been pursued. In contrast, the other minerals sector, being outside of the centralized policies of the Chamber, if anything significantly raised their nonwhite labor use in this same period, the coefficient on the post-1974 dummy being positive for that sector.[14]

As noted in connection with (10), if additional domestic labor could be hired without raising the wage or lowering productivity, it could not be argued that a policy of reliance on domestic labor is costly to the industry. However, the production function estimates for gold in Table 3 indeed show a significant decline in productivity after 1974 as the labor force becomes increasingly South African. Far from increasing productivity, as intended with a policy of diminished turnover, the replacement of skilled foreign workers by inexperienced South African novices proved costly. It was also costly in terms of wages, for the estimated supply equation for South African labor (in Table 2) again displays a significant upward slope with respect to mine wage. That estimate also brings out the role of the aspiration for a far better paying wage in industry in promoting reluctance to commit to the indignities of mine work, at least until 1978. Thereafter, compulsory relocation in the homelands has resulted in an excess supply even of South African workers for the mines.

C. *White and Nonwhite Labor*

Thus, the movement toward localization of the nonwhite labor force adversely affected productivity in gold, though some mining houses clearly hoped stabilization would enhance skill levels. On the other hand, in the diamond sector, erosion of the job color bar after 1964 and reduced turnover of the nonwhite labor force proved more successful. From the production functions for diamonds in Table 3, it is seen that the productivity of nonwhite labor is estimated to have a significant upward trend after 1964. This is also supported by the upward trend in the demand for nonwhite labor, at least in the quadratic case in Table 1.[15]

But concurrent with the rising productivity and demand for nonwhites in diamonds was a decline in both for whites. The coefficient on the white labor $\times$ time interaction after 1964 is significantly negative in the production functions of Table 1, as is the corresponding time effect in the white labor input equation in Table 4. It seems the erosion of the job color bar indeed had precisely the effects on white labor the Mine Workers' Union has always feared. Thus, in the production functions, the coefficient on the nonwhite $\times$ white labor interaction after 1964 proves negative though significantly so only in the quadratic case. With this, the nonwhite labor input equations are quite consistent, but not those for whites. It seems at least overt displacement of white labor by nonwhite hiring may have been successfully resisted despite productivity implications.

Indeed, from both Tables 1 and 4, the hiring of white and nonwhite labor is estimated to have been complementary across

[14] No estimate of the translog nonwhite labor input equation is reported for the other minerals sector owing to a high level of collinearity encountered.

[15] The sale price of diamonds is quite volatile. The estimated labor input equations therefore deflate wage by price lagged one period.

TABLE 4—WHITE LABOR INPUT EQUATIONS

	Quadratic				Translog			
	Gold	Coal	Diamonds	Other	Gold	Coal	Diamonds	Other
Intercept	10121	419	959	1966	4.31	−4.82	−.469	.068
	(1.74)	(0.38)	(2.27)	(1.02)	(2.84)	(3.36)	(0.41)	(0.06)
Wage Term	−22.3	−1.14	−.008	−77.7	−.410	−.870	−.0001	−.0004
	(4.24)	(5.11)	(0.02)	(1.45)	(2.37)	(2.97)	(0.32)	(0.51)
Nonwhite Labor,	.011	.086	.006	.063	.105	.818	.169	.534
	(0.83)	(5.89)	(0.43)	(9.32)	(0.85)	(5.52)	(2.93)	(6.79)
Equipment	.004	−.005	−.010	−.007	.165	−.019	−.023	.074
	(0.52)	(2.15)	(1.91)	(0.73)	(2.38)	(0.45)	(0.40)	(0.77)
Fixed Capital	−.006	.005	.025	.023	−.308	−.082	.403	.152
	(1.18)	(0.99)	(6.92)	(2.51)	(3.75)	(0.98)	(6.97)	(1.92)
Time, Post-1964			−111				−1.28	
			(8.17)				(7.07)	
Nonwhite Labor, Post-1964			.135				.387	
			(8.22)				(7.00)	
Lagged Dependent Variable	.799	.511	.694	.228	.671	.671	.418	.046
	(5.86)	(5.13)	(6.44)	(1.86)	(5.55)	(4.60)	(3.51)	(0.33)
Rho	.316	.039	−.521	.866	.420	.140	−.319	.686
Standard Error of *Rho*	.188	.186	.165	.148	.201	.130	.180	.150
Number of Observations	33	25	33	33	33	25	33	33

all sectors.[16] With tight maintenance of job delineation between races, additional nonwhite labor has required extra whites in supervisory and skilled posts rather than displacing them. It is also interesting to note that white labor has been a quasi-fixed factor for all sectors, in the sense that the flexible accelerator terms are positive, reflecting a significant difficulty in rapidly adjusting the white labor contingent. In contrast, as noted earlier, the centralized recruiting and prevention of worker organization have permitted far greater flexibility for the mines in adapting the nonwhite labor force numbers. To this, there is an exception in diamonds, where the lagged dependent variable term for nonwhites proves positive though lower than for whites. It seems that one cost to stabilization in diamonds has been a reduced ability or willingness rapidly to transform employment of nonwhites.

[16] The white labor equations for the coal sector omit the years prior to 1956. During 1955, the Mine Workers' Union demanded and received a sharp pay increase in coal. The bargaining position of the union was strengthened by a perceived white labor shortage in coal prior to this and the industry reponded by intensified training and recruiting of whites, with significant increased immigration of European trainees. By 1956, the Chamber's *Annual Report* (1957) notes an easing in the shortage of European personnel.

D. *Monopoly in Diamonds and Wage Discrimination*

No support is found in the data for either a hypothesis of concern for monopoly in labor application to diamonds or of wage discrimination against nonwhites elsewhere.

The estimated coefficients on the wage terms in the labor input equations for diamonds embody three components: the expression involving the elasticity of demand for diamonds, δ; the productivity elements, $\beta\beta$ or ll; and the flexible accelerator terms. The critical expression $d=1/(1+1/\delta)$ may, however, be isolated by imposing nonlinear cross-equation and within-equation restrictions on the other two components. Note that d approaches one from above as δ

tends toward minus infinity. A null hypothesis of behavior as if product demand were perfectly elastic may therefore be tested by means of a Gallant-Jorgenson *chi*-squared test on the change in sum of squared residuals when $d=1$ is imposed. In no case—white or nonwhite, quadratic or translog—does the estimate of d prove significantly greater than one. At least from this simple specification there is thus no evidence to suggest labor hiring has been restricted in diamonds for fear of depressing the world price of diamonds. Certainly DeBeers does limit diamond sales at times when there is downward pressure on prices, but it is interesting to note that the correlation coefficient for the first differences in physical sales and production of diamonds in South Africa is only .61 over the sample period. It seems at least temporarily restricted sales for price support result in increased stock piling rather than reduced production. To some extent this may be precisely because of the difficulties of quickly adjusting the labor force as emphasized by our estimates. In addition, DeBeers has the option of reducing their scale of production outside of South Africa and may have been more reluctant to make such adjustments within their South African operations.

Last, consider the possibility of wage discrimination against nonwhites in addition to the job barriers embodied in the production functions. One way to model such potential discrimination is to envision the mines hiring as if this imposed some implicit psychic cost, c, for each additional nonwhite hired. Thus, c would be added to the right-hand side of equations (3) and (3′a). To the estimated nonwhite equation in (10) this would add a term $c/(p\cdot\beta\beta)$, (multiplied by b/q in the translog case). If the nonwhite labor equations for gold, coal, and others are reestimated including such a term, the resulting value of $c/\beta\beta$ is significantly negative only in the quadratic case for coal, and even there fails a 5 percent one-tail test. There is thus no systematic evidence of wage discrimination against nonwhites from this simple test, though obviously job discrimination is quite overt under the apartheid structure.

V. Some Closing Remarks

Despite the popular notion that much of southern Africa formed a vast labor reserve for the South African mines—a traditional surplus labor economy—the supply curve on which the industry operated has been shown not only to be upward sloping with respect to mine wages of nonwhites, but indeed inelastic. In the period 1946–74, the Chamber of Mines took significant advantage of this —principally through the dominant gold-mining sector—to exploit its monopsony power. By the mid-1970's, rising mineral prices permitted sharp wage increases motivated by a desire to stabilize the labor force (in the sense of reducing turnover) and by vehement public criticism of the "starvation wages" previously paid. The rising wages attracted more South African workers and on their supply curve the industry has continued to act as a monopsonist. In this later period, foreign labor has been in excess supply and a significant preference not to rely on foreign labor has been demonstrated. Such a preference is not irrational from the mining industry's perspective, given the political uncertainties of the region as, for example, FRELIMO displaced the cooperative Portuguese in Mozambique and Malawi's labor was temporarily withdrawn entirely. But stabilization of the South African labor force has not, thus far, resulted in significant productivity gains, rather the foregoing results suggest diminished productivity as South African novices increasingly displaced experienced foreign miners. Moreover, black mine workers have now succeeded in achieving a greater degree of organization, previously thwarted by the mines precisely by maintaining diversified options on foreign labor.

The job color bar, despite limited liberalization of job tasks, has been protected by the apartheid laws partly under instigation from the Mine Workers' Union. Thus, maintenance of this bar has meant that black and white labor have acted as complementary factors, and barring nonwhites from skilled positions helped to support the twenty fold white-nonwhite earnings ratio as

of 1970. To what extent the mining houses acquiesced in this legislated job discrimination remains unclear. Certainly the apartheid structure has imposed costs on the mining industry, some negotiated erosion of job barriers has occurred, and no evidence of wage discrimination by the mines is discerned, superimposed on the obvious job discrimination.

To much of this the diamond sector is an exception. Pursuing an independent wage policy since the mid-1960's, the diamond sector has stabilized its nonwhite labor force, eroded the skill barriers, and largely dispensed with the mine compound living quarters. Both the particular need for security and the personal involvement of Harry Oppenheimer are commonly cited as instigating factors in this transition. The result has been a significant rise in nonwhite labor productivity and a simultaneous decline in productivity of whites, though overt displacement of white miners seems to have been successfully resisted. An additional consequence for the diamond sector is an estimated flexible accelerator on nonwhite labor input more in line with that on white labor throughout the industry, reflecting the slower response in adjusting any stabilized labor force. Elsewhere the industry has shown greater swiftness in adjusting, via the recruitment network, its nonwhite labor force as demand conditions shifted—far more swiftness than permitted to labor supply bound by irreversible contracts under the master and servants acts.

Since 1979, even South African labor has been in excess supply following an outward shift in the supply curve induced by compulsory population relocation in the homelands. Today, the Chamber is in consequence split between these mining houses that wish to raise nonwhite wages further and those that argue for real wage reduction, with both realizing lower wages would stimulate more employment. To go their separate ways would, according to interviews with industrialists, effectively break the Chamber, and the Chamber is vital not only for labor purposes but for such activities as lobbying for tax positions. That the diamond sector could pursue its own wage policy while remaining a Chamber member for other purposes is attributable to the negligible size of its labor force. Government has generally encouraged the recent wage increases, but while concerned with the risks inherent in mass unemployment in the homelands, they would prefer to see employment created around the peripheries of the homelands rather than in mines located too close to the exclusive white environs.

The future of this labor market—historically of major importance to several of the labor supplying countries for employment, foreign exchange, and income—is very uncertain. Even if the South African internal political structure remains intact for some years, much will depend on future mineral prices. Recently, the value of diamonds and gold have substantially declined. From 1979 onwards, TEBA has maintained a target of 60 percent local, 40 percent foreign, labor. Some observers predict the foreign component will decline. If so, Lesotho is likely not to experience substantial cuts, for as with the homelands there is fear of creating pockets of mass unemployment and poverty within the frontiers of South Africa. The proportions in which the other labor sending countries would be cut would probably depend on their political conduct vis-à-vis South Africa. Indeed, it will be interesting to see if there may even be revival in Mozambique recruitment, now that a new, intergovernmental understanding has been signed with South Africa.

REFERENCES

Berndt, Ernst K. et al., "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement*, Fall 1974, *4*, 653–65.

Breytenbach, W. J., *Migratory Labour Arrangements in Southern Africa*, rev. ed., Pretoria: Africa Institute of South Africa, 1979.

Daniel, Philip, *Africanisation, Nationalisation and Inequality*, Cambridge: Cambridge University Press, 1979.

First, Ruth et al., *The Mozambican Miner*,

Maputo: Centro de Estudos Africanos, Universidade Eduardo Mondlane, 1977.

Fuss, Melvyn, McFadden, Daniel and Mundlak, Yair, "A Survey of Functional Forms in the Economic Analysis of Production," in M. Fuss and D. McFadden, eds., *Production Functions*, Vol. I, Amsterdam: North-Holland, 1978.

Gallant, A. Ronald and Jorgenson, Dale W., "Statistical Inference for a System of Simultaneous, Non-linear, Implicit Equations in the Context of Instrumental Variable Estimation," *Journal of Econometrics*, October-December 1979, *11*, 275–302.

Lipton, Merle, "Men of Two Worlds: Migrant Labour in South Africa," *Optima*, November 1980, *29*, 71–201.

Lucas, Robert E. B., "Emigration, Employment and Accumulation: The Miners of Southern Africa," Discussion Paper No. 4, Migration and Development Program, Harvard University, December 1983.

Nadiri, M. Ishaq and Rosen, Sherwin, *A Disequilibrium Model of Demand for Factors of Production*, NBER, New York: Columbia University Press, 1973.

Wilson, Francis, *Labour in the South African Gold Mines 1911–1969*, Cambridge: Cambridge University Press, 1972.

Yoshikawa, Hiroshi, "On the 'q' Theory of Investment," *American Economic Review*, September 1980, *70*, 739–43.

Transvaal and Orange Free State Chamber of Mines, *Annual Report*, Johannesburg, 1957.

Competition and Unanimity Revisited, Again

By CLEMENT G. KROUSE*

Conditions necessary and sufficient for shareholders to express unanimity with respect to firm decisions have been recently developed by Harry DeAngelo (1981). In a model economy where firms are properly "small relative to the market," Louis Makowski (1983) has subsequently found DeAngelo's requirement that existing securities span the opportunity set to be unnecessary for unanimity. At issue in this difference is what is meant by "reasonable" necessity in propositions relating to model economies: is Makowski's "small relative to the market" condition of interest, or is it, like the Giffen good, only a curiousity? While Makowski's case initially appears implausible, I shall here argue that the specification of an economy with "equivalently complete" markets is a reasonable instance of his analysis and qualifies as an interesting complement to the general propositions of DeAngelo.[1]

In Section I some basic notation and the structure of the model economy are set out. Following that, the distinction between the unanimity proofs of DeAngelo and Makowski is drawn and the issue of reasonableness in necessary conditions is addressed. Conditions under which equivalently complete markets arise are then described, and the reasonableness of these are considered along with their implications for unanimity.

I. The Model Economy

I consider the usual two-period economy. In the first (current) period all things are known with certainty. Events at the second (future) period, however, are uncertain and characterized by $s=1,2,\ldots S$ states of nature. The individuals of the economy exhibit diversity, being restricted only by the usual assumptions of utility-maximizing behavior with increasing, strictly quasi-concave, suitably continuous utilities defined on state-contingent rates of return.[2] Individuals choose portfolios of securities, with portfolio fractions summing to one.[3] A *security program* Z is a $J\times S$ state space tableau of returns offered by some collection of $j=1,2,\ldots J$ securities, $z^j=[z_1^j z_2^j \ldots z_S^j]$ being the rates of returns paid by security j over the list of states. To avoid unnecessary complexity, I momentarily assume that each firm issues just one security (equity) so that j indexes both firms and securities. Let $v=[v^1\, v^2 \ldots v^J]$ indicate equilibrium security prices in trading at the first period.[4]

When securities trade without restriction on short sales, then $\text{rank}(Z)=K$ and a $(k-1)$ dimensional hyperplane is the subspace spanned by all portfolios of the existing securities. I use $\Omega(Z)$ to indicate that subspace and call this the *portfolio opportunity set*. A collection of linearly independent security portfolios which are spanning vectors for $\Omega(Z)$ defines a set of security markets. A second security program Z^*, and associated portfolio opportunities $\Omega(Z^*)$, is unambiguously said to offer new markets or, equivalently, to provide increased trading opportunities relative to Z when $\Omega(Z)$ is a proper subset of $\Omega(Z^*)$.

Taking current consumption as numeraire, a *sustaining state claims price system* $\pi=$

*Professor of Economics, University of California, Santa Barbara, CA 93106. I thank Harry DeAngelo, Louis Makowski, and John Riley for helpful comments.

[1] Roughly, security markets are said to be equivalently complete when a nontrivial set of mutual funds spans the set of consumption portfolios optimal for all individuals in the economy. A more detailed definition is given below.

[2] When defining utility on rates of return, utility functions reflect not only preferences but also wealth (see my 1984 study, ch. 5).

[3] While Makowski generally places his analysis in the firm's product market, I restrict attention here to the capital market, which both he and DeAngelo consider.

[4] It is assumed throughout that markets are perfect in the sense described by DeAngelo (pp. 21–22).

$[\pi_1\pi_2 \dots \pi_S]$ is such that

$$(1) \qquad 1 = \Sigma_s \pi_s z_s^j, \quad \text{or} \quad v_j = \Sigma_s \pi_s z_s^j,$$

gives the market prices v^j for each security j, with each z_s^j being the dollar payoff associated with the z_s^j rate of return. When each individual's optimality occurs at a stationarity point, the personal rates of substitution of state claims for current consumption qualify as a system of sustaining prices and, from the obvious algebra, so do all convex combinations of these systems. For the price system of some specific individual i, I shall use the notation π^i, referring to this as i's *demand prices*. The capital market is said to be *complete* when $K = S$, and from equations (1) it is then seen that the sustaining price system π is unique in equilibrium. When, alternatively, $K < S$ the capital market is *incomplete* and an infinite number of sustaining price systems exist, with only those portfolios in $\Omega(Z)$ being priced identically by every such system. Conversely, the S-vector $w \notin \Omega(Z)$—we say that w is not spanned by the existing securities—will in general be priced differently by different sustaining price systems.[5]

Consider next a specific security program Z with associated opportunity set $\Omega(Z)$. Let $K < S$ so that markets are incomplete. A planned investment by some firm k, all other firms holding their outputs constant, alters the state contingent return to k's security in the future period, say to $(z^k + \Delta z^k)$. When this planned security lies "outside" the existing market, $(z^k + \Delta z^k) \notin \Omega(Z)$, different pre-investment sustaining state claims price systems will not generally yield the same value to post-investment portfolios. Specifically, for individuals i and n the pre-investment demand prices will be such that

$$\Sigma_s \pi_s^i \left(z_s^k + \Delta z_s^k \right) \neq \Sigma_s \pi_s^n \left(z_s^k + \Delta z_s^k \right).$$

When these valuations differ across shareholders with sufficiently heterogeneous preferences and/or endowments, some will find the planned investment value to be in excess of the initial outlay while others will not. In this case, unanimity does not obtain. And, when this occurs, "value-maximizing" firm decisions will be ill-defined, for whose prices are to be used to determine value?

II. Enter DeAngelo and Makowski

This background provides the motivation for DeAngelo's dual conditions for unanimity. His Theorem 1 implies that every set of pre-investment demand prices provide an unambiguous basis for valuing the planned investment when: 1) (opportunity set spanning) $z^k + \Delta z^k$ lies in $\Omega(Z)$; and 2) (scale) the change Δz^k is "small" relative to the total supply of returns in each state. Jointly, these conditions imply competitive conjectures: the scale condition is usual and the spanning condition makes clear the "scale of what."[6] When these conditions occur simultaneously, each shareholder perceives his demand prices to be unaffected. Furthermore, at fixed-claim prices, the indirect utility functions of the individuals will be increasing in wealth, which means that value-maximizing firm decisions will receive unanimous shareholder support.

In his analysis, Makowski has argued that opportunity set spanning is not needed for unanimity. Just as DeAngelo, he would have individuals use their existing state claim prices to value planned investment decisions, and this he finds to be possible when firms are properly small relative to the market. When will such "smallness" occur?

Makowski provides several illustrative examples. Using the above notations, we can quickly see to the core of these. First, assume that firm k's intentions are to continue issuing security z^k and, in addition, offer a new Δz^k security not spanned by securities of the pre-investment security program Z; that is, $\Delta z^k \notin \Omega(Z)$. Let Z^* indicate the opportun-

[5] See, for example, Phillip Dybvig and Stephen Ross (1982) or my 1984 study, ch. 4.

[6] There would in general have to be offsetting changes by other firms to keep the social total of claims exactly the same in each state and thus assure *actual* price taking. DeAngelo covers this matter in detail. See also my 1978 article.

ity set with Δz^k, noting in turn that $\Omega(Z)$ is a proper subset of $\Omega(Z^*)$ so that new trading opportunities are created. Now rank the individuals in the economy by the values they place on the proposed security Δz^k at the existing, pre-investment demand prices. Consider specifically the set of individuals whose demand prices provide the (single) greatest shadow value to Δz^k. Allocate Δz^k only to the members of this set. At this point, Makowski would have us think of some combination of a small Δz^k and a large number of individuals in this set so that each receives an insignificantly small increment to *ex ante* consumption: small thus means that the allocation does not (significantly) affect any such individual's optimal portfolio, utility, and thus marginal valuation of state claims (and, therefore, demand prices for Δz^k). In short, the new trading opportunities are not valued by individuals; in Makowski's terminology, "no surplus" is created. For the smallness requirement to have application in a finite economy, we thus see requires a quite special collection of individuals and/or limitations on possible productions.

III. Unanimity and Equivalently Complete Markets

What conditions are sufficient for Makowski's "no surplus?" A simplified example will suggest one interesting case. For this it is convenient to restrict attention to a pure capital structure decision—the relevance of the analysis to the investment (production) decision will be obvious in a moment. Think specifically of the case with three firms and four states of nature with rates of return given by the security program Z of Table 1. The three securities are linearly independent, spanning a 3-dimensional subspace of the 4-dimensional state space (the security market is incomplete). It can be easily verified that a system of sustaining state claims prices for Z is $\pi_1 = 0.3$, $\pi_2 = 0.2$, $\pi_3 = 0.3$, and $\pi_4 = 0.1$.

It is useful to denote the set of portfolios which are optimal for all individuals in the model economy by $\Omega^0(Z)$ and call this the *efficient set*: note that $\Omega^0(Z)$ is a (generally proper) subset of $\Omega(Z)$. Now suppose, illustratively, that individuals maximize expected utility with quadratic elementary utility functions. From the early work of Harry Markowitz (1952) and James Tobin (1958) in this case, it is known that each individual chooses an optimal portfolio as a linear combination of the riskless mutual fund $R = z^1$ and the all-risky aggregate (or market) mutual fund $M = (1.6/3.1)z^2 + (1.5/3.1)z^3$. That is, portfolios of the two funds

$$R = [1.111 \quad 1.111 \quad 1.111 \quad 1.111]$$

$$M = [0.969 \quad 1.612 \quad 0.967 \quad 0.967]$$

yield all optimal plans. We say that M and R span the efficient set: $\Omega^0(Z)$ is a plane in the state space.

With this as background, now suppose that firm 3 undertakes a capital structure decision: it issues risky debt with a promised payment of \$1.5, $z_s^{3d} = \min[z_s^3, 1.5]$, and equity, $z_s^{3e} = z_s^3 - z_s^{3d}$, instead of its single security paying z_s^3 in each state s. To avoid problems of expropriation and wealth effects, assume that *ex ante* ownership shares in z^{3e} and z^{3d} are the same as in z^3. The security program Z^* that results is given in Table 2. Notice that $z^3 = z^{3d} + z^{3e}$, for it is assumed that the capital structure change is without drains or subsidies.

Having not changed individual utilities, the existence of the riskless asset, or the aggregate total of risky state claims, the effi-

TABLE 1—SECURITY PROGRAM Z

Securities	States			
z^1	1.111	1.111	1.111	1.111
z^2	1.250	1.875	0.625	0.625
z^3	0.667	1.333	1.333	1.333

TABLE 2—SECURITY PROGRAM Z*

Securities	States			
z^1	1.111	1.111	1.111	1.111
z^2	1.250	1.875	0.625	0.625
z^{3d}	0.833	1.250	1.250	1.250
z^{3e}	0.000	1.667	1.667	1.667

cient set does not change with the capital structure change: $\Omega^0(Z)=\Omega^0(Z^*)$. Thus, even though there are new markets and new trading opportunities—$\Omega(Z)$ is a proper subset of $\Omega(Z^*)$—*optimal* plans for individuals can again be written as a portfolio of the same two mutual funds R and M.[7] And, with no changes in equilibrium there will be no changes in individual demand prices. In Makowski's terminology, we can say that no surplus is created by the new markets offered with Z^* relative to Z.

By induction from the above case it is seen that when more than two mutual funds span the efficient set, and these are not altered when changing from an original program Z to some new Z^*, then the no surplus condition continues to hold. That is, as long as neither wealth positions nor the efficient set are altered with changes in the portfolio opportunity set, new trading opportunities do not change individual consumption plans or demand prices.

Notice in the example above that the change from Z to Z^* completed the market: the four securities of the revised program Z^* have linearly independent returns. Thus, insofar as individuals are concerned, the original market Z provides all the *relevant* consumption possibilities of a complete market, and, in equilibrium with either Z or Z^* (all other things constant), the allocation of risk in the cross section of individuals is identical. More generally, let Z and Z^* be two security programs where $\text{rank}(Z^*)=S$, $\text{rank}(Z)=K$, and $K<S$. Z^* represents a complete market, but Z does not, which means that $\Omega(Z)$ is a proper subset of $\Omega(Z^*)$. We say that the markets associated with the security program Z are *equivalently complete* if $\Omega^0(Z)=\Omega^0(Z^*)$.

Suppose that the number of individuals I is less than the number of securities. Then, clearly at most I mutual funds will span the efficient set. In large economies, this case is of little significance, but it does raise the question of when there will be a small number of such spanning mutual funds. In this regard it is possible to limit the class of individuals to some subset of those who are risk-averse, expected-utility maximizers, limit the joint distribution on returns of the securities, or use some combination of these two limitations.

IV. κ-Fund Spanning

When a riskless asset exists, security return distributions are not restricted (except that have appropriately finite moments), and we make the usual assumption that individuals are risk averse with homogeneous probability beliefs, then two-fund spanning of $\Omega^0(Z)$ occurs if and only if each elementary utility function exhibits linear risk tolerance with a common "slope" parameter[8] (see Mark Rubinstein, 1974; and Michael Brennan and Alan Kraus, 1976). By specific choice of the intercept and slope parameters in the expression of linear risk tolerance, these utility functions can be made to possess increasing, constant, or decreasing absolute or relative risk aversion. This includes all of the commonly used quadratic, logarithmic, exponential, and power elementary utilities. In this respect the limitation to linear risk tolerance does not appear to be overly restrictive. On the other hand, the requirement that the slope parameter be common means that individuals in the economy can differ only to the extent allowed by the intercept parameter and their investment wealth. The commonality of slope requires that each and every utility function belong to a single subclass. For example, when one individual has logarithmic utility, then so must all others, and this holds also for exponential utility. Further, if one individual has utility represented by a power function, then all individuals must have power functions of the same power. Finally, note too that the analysis has assumed homogeneity of probability esti-

[7] While M has the same pattern of state-contingent returns, note, however, that it is made up somewhat differently as $M=[(1.6/3.1)z^2+(1.2/3.1)z^{3d}+(0.3/3.1)z^{3e}]$.

[8] Let $r(w)$ be the Arrow-Pratt measure of absolute risk aversion. Then $T(w)=1/r(w)$ is the risk tolerance function and $T(w)=\alpha+\beta w$ indicates utility that is linear in risk tolerance.

mates across individuals, so that on closer inspections individuals cannot differ by "very much" at all if two fund spanning is to occur by specification of individual characteristics.

While the class of elementary utilities yielding $\kappa > 2$ fund spanning are somewhat less restrictive, on close inspection David Cass and Joseph Stiglitz have concluded that "...[efficient set] spanning turns out to be a property limited to a very, very few utility functions" (1970, p. 144). In reaction to this, Stephen Ross (1978) has studied the alternative case, conditions on the joint distribution of security returns necessary and sufficient for efficient set spanning. Roughly, limiting individuals only to expected utility maximization, he shows that κ mutual funds span $\Omega^0(Z)$ if and only if there exist constants a_{jk} $(k=1,2,\dots\kappa)$ such that, for all j,

$$z_j = \sum_k^{\kappa} a_{jk} m^k + \varepsilon_j,$$

where the m^k are the spanning mutual funds and the ε_j are mean zero noise with respect to the mutual fund returns. This restriction is satisfied for $\kappa = 2$ when security returns are joint normally distributed, or more generally are in the class of 2-moment stable distributions, or when the joint distribution is symmetric. κ-moment stable distributions assure κ-fund spanning of $\Omega^0(Z)$, but not much is known about specific conditions for higher order spanning other than that.

When interpreting the m^k as common factors and imposing mild restrictions on the ε_j, Ross has, in addition, derived an asymptotic spanning theorem. Specifically, he shows that as the number of securities become large, the κ common factors become perfectly correlated with κ mutual funds independent of any other distributional assumptions on the factors or noise terms. As empirical studies of security returns have generally found fewer than five statistically significant common factors, and available securities number in the thousands, this approximation result appears to have a reasonable empirical basis.[9] Efficient set spanning with a few "mutual funds" thus appears to be a case of interest.

[9]See Richard Roll and Ross (1980).

V. Production Decisions

To this point I have restricted attention to capital structure changes. In perfect markets without drains and subsidies these "packaging" changes can affect trading opportunities —they can alter the opportunity set—but they do not alter the firm's overall output or, therefore, the aggregate (social total) of state claims. In contrast, changes in the firm's output plan (production) generally affects trading opportunities in the same way as a change in capital structure, *and* it also generally alters the aggregate total of state claims. For competitive production, it is thus necessary that any new trading opportunities that are created not provide value by altering the efficient set as described above *and* that the additions to aggregate claims in every state be insignificantly small in the usual sense.

VI. In Conclusion

These observations suggest the following corollary to DeAngelo's basic theorem.

COROLLARY: *Let conditions for equivalently complete markets occur. Then, if the ex ante shareholders of a firm also have competitive conjectures with respect to scale, net value maximization will be unanimously supported as a firm objective.*

As a final remark, let me note that it is usual to think that the number of states of nature is very large and the number of securities is relatively small in comparison. Markets are as a result thought to be incomplete and this has precipitated the concern for differences between full and constrained Pareto optimality of risk allocation and, more fundamentally, the concern for unanimity and an appropriate firm objective. Today, I think most economists involved with capital markets are less sure of the relevant incompleteness of such markets. That is for several reasons. We now understand that the proper question is whether or not the efficient set of

plans (and not the entire consumption set) is spanned with changes in firm decisions. Moreover, the option feature of stocks and bonds and the compound option nature of the securities derived from them (puts, calls, warrants, etc.) are now known to complete markets in ways that do not imply large (incremental) information and transaction costs. And, finally, it is now being realized that under not too restrictive conditions, somewhat coarser partitions of states than heretofore thought are relevant (see, for example, my 1984 study, ch. 10).

REFERENCES

Brennan, Michael and Kraus, Alan, "The Geometry of Separation and Myopia," *Journal of Financial and Quantitative Analysis*, June 1976, *11*, 171–93.

Cass, David and Stiglitz, Joseph E., "The Structure of Investor Preferences and Asset Returns, and Separability in Portfolio Allocation," *Journal of Economic Theory*, June 1970, *2*, 122–60.

DeAngelo, Henry, "Competition and Unanimity," *American Economic Review*, March 1981, *71*, 18–27.

Dybvig, Philip and Ross, Stephen, "Efficient Portfolio Sets," *Econometrica*, October 1982, *50*, 1525–46.

Krouse, Clement, "The Optimality of Risk Allocation: A Synthesis." *Southern Economic Journal*, April 1978, *22*, 762–77.

______, "Capital Markets and Prices: Valuing Uncertain Income Streams," manuscript, University of California-Santa Barbara, 1984.

Makowski, Louis, "Competition and Unanimity Revisited," *American Economic Review*, June 1983, *73*, 329–39.

Markowitz, Harry, "Portfolio Selection," *Journal of Finance*, March 1952, *29*, 77–91.

Pye, Gordon, "Portfolio Selection and Security Prices," *Review of Economics and Statistics*, February 1967, *49*, 111–15.

Roll, Richard and Ross, Stephen, "An Empirical Investigation of Arbitrage Pricing Theory," *Journal of Finance*, October 1980, *5*, 1073–103.

Ross, Stephen, "Mutual Fund Separation in Financial Theory—The Separating Distributions," *Journal of Economic Theory*, June 1978, *17*, 254–86.

Rubinstein, Mark, "An Aggregation Theorem for Security Markets," *Journal of Financial Economics*, September 1974, *1*, 225–44.

Tobin, James, "Liquidity Preference as a Behavior Towards Risk," *Review of Economic Studies*, February 1958, *25*, 65–86.

Wilson, Robert, "The Theory of Syndicates," *Econometrica*, January 1968, *26*, 119–32.

Inflation, Real Balances, Output, and Real Stock Returns

By JASON BENDERLY AND BURTON ZWICK*

Over the 1954–81 period, as well as earlier in the century, a negative simple correlation existed between the real return on common stocks and the rate of inflation in the United States. This evidence of an inverse relation contrasts with the traditional theory that the rate of return on equities should be invariant with respect to nominal variables such as inflation.

Eugene Fama (1981) concludes that the negative simple correlation between real stock returns and inflation is spurious because it is the result of two structural relationships: a positive relation between current real stock returns and expected output growth, and a negative relation between expected output growth and current inflation. Fama first concludes that in an efficient market, expected output or earnings growth is the primary determinant of stock returns. He then explains the negative relation between inflation and expected output growth by particular models of money demand and the equation of exchange in which causality runs from expected (future) output growth to current inflation, rather than from current inflation to future output.

This paper extends Fama's efforts to show that the negative simple correlation between the rate of inflation and real stock returns is spurious. In Section I, we focus on the positive relation between current real stock returns and expected (future) output growth. We present estimates of stock return equations for periods other than Fama's 1954–76 sample. In Section II, we focus on the negative relation between current inflation and future output growth. We present an alternative explanation of the inverse inflation-future output relation based on a real balance model of unemployment and output originally presented by Jerome Stein (1982) and in our earlier article (1985). In contrast to Fama's analysis, we view the causality as running from current inflation to future output growth, rather than from future output growth to current inflation. We show how Fama's inflation estimates are more plausibly interpreted as a transformation of our estimated output model. We then integrate the stock returns-future output results from Section I and the output-inflation results from Section II into an overall relation among real stock returns, future output growth, and current inflation.

I. Inflation, Output, and Real Stock Returns

Fama argued that, in an efficient and forward-looking market, real stock returns should reflect expectations only about real variables, such as growth in output or production. Any inverse relation between real stock returns and inflation must reflect an inverse relation between inflation and future output or production growth.[1]

To test for an effect of inflation on real stock returns that is independent of future output growth, we follow Fama and regress real stock returns (RS_t) against next year's output growth ($\%Q_{t+1}$) and this year's inflation rate ($\%P_t$).[2] Using ordinary least squares,

*Kidder, Peabody and Company, 10 Hanover Square, New York, NY 10005. We are grateful to Michael Hamburger for extensive discussions, and appreciate the comments of Ira Kaminow, Hugh Neuburger, and Richard Roll, the research assistance of Gary Bigg, and the typing of Michele Pirone.

[1]Reuben Kessel (1956), Zvi Bodie (1976), Jeffrey Jaffe and Gerschon Mandelker (1976), Charles Nelson (1976), and Fama and G. William Schwert (1977), among others, report significant simple correlations between real stock returns and the contemporaneous inflation rate. These studies were in part the motivation for Fama's analysis.

[2]*RS* is an *ex post* rather than an *ex ante* measure of returns. Most analysts interpret the effect of expected output growth on *RS* as the result of output's effect on expected earnings and dividends. Robert Shiller (1984) suggests that the effect of output growth on *RS* also reflects the effect of output growth on the discount rate. The supply of shares being relatively fixed, increased

parameters were estimated with annual data from 1954–76 (Fama's original period of estimation), 1956–76 (because of unusual stock return behavior in 1954 and 1955 as discussed below), and 1956–81 (to pick up the latest period for which data on $\%Q_{t+1}$ is available). Because of first-order autocorrelation, the 1954–76 estimates were reestimated using generalized least squares. RS is measured using the Ibbotson-Sinquefeld data base, Q is measured by real GNP, and P is measured by the deflator for personal consumption expenditures.[3] Following Fama, we measured stock returns using year end (December) data, output growth using annual average data, and inflation using year-end (December) data. That is, RS_t is the real return from December of year $t-1$ to December of year t, $\%Q_{t+1}$ is growth in output from year t to year $t+1$, and $\%P_t$ is the growth of the price level from December of year $t-1$ to December of year t (t-statistics are shown in parentheses).

1954–76

$$\underset{}{RS_t} = \underset{(-.42)}{-.042} + \underset{(5.28)}{6.50\%Q_{t+1}} \underset{(-1.71)}{-2.91\%P_t}$$

$$\bar{R}^2 = .71;\ D\text{-}W = 2.01,\ S.E. = .118,\ \underset{(2.40)}{\rho = .48}$$

1956–76

$$RS_t = \underset{(-1.50)}{-.116} + \underset{(5.17)}{6.04\%Q_{t+1}} \underset{(-.88)}{-1.05\%P_t}$$

$$\bar{R}^2 = .70,\quad D\text{-}W = 1.70,\quad S.E. = .106$$

1956–81

$$RS_t = \underset{(-1.49)}{-.124} + \underset{(4.12)}{5.23\%Q_{t+1}} + \underset{(.17)}{.19\%P_t}$$

$$\bar{R}^2 = .47,\quad D\text{-}W = 1.71,\quad S.E. = .133.$$

For the 1954–76 period used by Fama, the coefficient of inflation is almost statistically significant ($t=-1.71$) and suggests an effect of inflation on real stock returns independent of the stable and statistically significant effect of future output. This is the counterpart to Fama's equation ($A6$) (p. 558), which regresses RS_t against $\%Q_{t+1}$ and measures of expected and unanticipated inflation. However, for 1956–76 and 1956–81, the output coefficient remains highly significant, but the inflation coefficient is no longer significantly different from zero. The reason that the 1954 and 1955 observations are sufficient to alter the inflation-stock return relation is apparent from the time-series of RS_t, $\%Q_{t+1}$, and $\%P_t$ presented in Table 1. The 1954 real stock return of 53 percent is the highest of the sample period while the inflation rate of $-.4$ percent is the lowest of the period. The 1955 return of 31 percent is the third highest of the period, while the inflation rate of 0.4 percent is second lowest of the period.

In addition to the sample periods just discussed, twelve window regressions 15 years in length were run over the 1956–81 sample period. For all 15-year intervals since 1956, the inflation coefficient is insignificant while the output term is highly significant. As already mentioned, the 1954 and 1955 observations were the major outliers of the post-Korean War period. Despite the divergent pattern of returns in 1954 and 1955, we made one final estimate—from 1954 (the initial year in Fama's sample) through 1981 (the most recent year for which $\%Q_{t+1}$ is available). The 1954–81 sample period estimate also shows a significant output coefficient [$6.25(t=3.71)$] and an insignificant coefficient on inflation [$-.99(t=-.64)$]. Even including the 1954 and 1955 observations, the independent effect of inflation on real stock returns observed in Fama's 1954–76 estimates disappears when the estimation period is extended through 1981.

output raises the demand for shares irrespective of any changes in expected earnings and dividends. Whatever the reason for the effect of expected output growth on RS, the measure RS should not be confused with an *ex ante* real return or any other measure assumed to affect capital investment.

[3]In measuring inflation and deflating the Ibbotson-Sinquefeld series of nominal stock returns, we used the personal consumption expenditures deflator rather than the Consumer Price Index because of factors—particularly the treatment of housing costs—that caused the *CPI* to overstate the rate of inflation in the late 1970's and 1980. None of our results are materially affected by the particular measure of inflation chosen.

TABLE 1—REAL STOCK RETURNS, OUTPUT GROWTH, AND INFLATION: 1954–81
(Shown in Percent)

	Real Returns (RS_t)	Output Growth ($\%Q_{t+1}$)	Inflation ($\%P_t$)
1954	53.0	6.7	−0.4
1955	31.2	2.1	0.4
1956	3.7	1.8	2.9
1957	−13.8	−0.4	3.0
1958	41.7	6.0	1.7
1959	10.5	2.1	1.5
1960	−1.3	2.6	1.8
1961	26.1	5.8	0.8
1962	−10.5	4.0	1.8
1963	21.2	5.3	1.6
1964	15.5	6.0	1.0
1965	10.2	6.0	2.3
1966	−13.3	2.7	3.2
1967	21.3	4.6	2.7
1968	6.8	2.8	4.3
1969	−13.5	−0.2	5.0
1970	−0.4	3.4	4.4
1971	10.5	5.7	3.8
1972	15.4	5.8	3.6
1973	−22.6	−0.6	7.9
1974	−37.3	−1.2	10.8
1975	31.2	5.4	6.0
1976	19.1	5.5	4.7
1977	−13.1	5.0	5.9
1978	−1.3	2.8	7.9
1979	8.6	−0.3	9.8
1980	22.2	2.6	10.2
1981	−12.2	−1.9	7.3

By our interpretation, these estimates show a consistent and statistically significant effect of next year's output on real stock returns. Given next year's output, the effect of reported inflation on real stock returns is not significant. This result appears consistently in estimates over all intervals excluding 1954 and 1955. And even when 1954 and 1955 are included, the significant effect of inflation disappears in estimates extending through 1981.

In his estimation of stock returns, Fama separated inflation into expected inflation ($EITB$) and unexpected inflation ($UITB$). The $EITB_{t-1}$, expected inflation from December of $t-1$ to December of t, is inflation as predicted from a model in which inflation in t is regressed against the Treasury bill rate at the end of $t-1$, and $UITB_{t-1}$ is the unexplained variance of inflation.[4]

Fama reports a statistically significant effect of $EITB_{t-1}$ ($t=-1.84$), given $\%Q_{t+1}$ and $UITB_{t-1}$, in estimates for 1954–1976. Only by including monetary base growth was Fama able to eliminate the effect of $EITB_{t-1}$. However, when we excluded 1954 and 1955, the t-statistic on $EITB_{t-1}$ falls to less than .5. Just as with reported inflation, these estimates show no significant effect of expected or unexpected inflation in estimates excluding 1954–1955. We then used the Livingston survey measure of expected inflation to partition actual inflation into expected and unexpected components. Using the Livingston measure, the coefficients of both expected and unexpected inflation are insignificant even when 1954–55 are included.[5]

We interpret the results of this section as very strong support for Fama's argument that the negative simple correlation between the level of inflation and real stock returns is the result of a second negative relation between inflation and future output growth. Our explanation of this negative relation is provided in Section II, Part A.[6]

[4] For a detailed discussion of the derivation of $EITB$, see Fama and Michael Gibbons (1982).

[5] Both of these measures of expected and unexpected inflation follow Fama and Livingston in using the CPI.

[6] Fama's analysis—and our analysis in Section I—focuses on an inverse relation between real stock returns and the *level* of the inflation rate. It does not relate to the argument of Martin Feldstein (1980) which links real returns to the *change* in the rate of inflation. (More precisely, Feldstein's theory relates returns to a change in expected rather than reported inflation.) His analysis is based on the nonneutrality of the tax system due to historical cost depreciation and the taxation of nominal rather than real investment returns. We regressed real returns against next year's output and the contemporaneous change in inflation ($\Delta\%P$). The change in inflation, in contrast to the level of inflation, was statistically significant and suggests that stock return models based on future output growth can be improved by including the contemporaneous change in inflation. Reflecting Feldstein's emphasis on expected rather than reported inflation, we replaced $\Delta\%P$ with first differences in the Livingston and Fama (Treasury bill rate) measures of expected inflation. Neither measure was significant. The inferior performance of these expected inflation measures suggests that 1) the significance of

II. The Real Balance Effect

The real balance effect, as originally formulated by Pigou and Patinkin, refers primarily if not solely to the direct effect of real money balances on private expenditures, particularly consumption. Subsequent elaboration, by monetarists and also neo-Keynesians such as Franco Modigliani (1971) in his formulation of the *FMP* model, has extended the real balance effect to include not only the direct effect of real balances on the demand for goods, but also the effect of real balances on asset prices, which in turn affect the aggregate demand for goods. Particularly when extended to include the transmission of monetary changes through the financial sector, the real balance effect occupies an important role in almost all monetarist and neo-Keynesian analyses of aggregate demand. Our analysis of output in Section II, Part A, and stock returns in Section II, Part B, is based on the theory of a real balance effect.[7]

$\Delta\%P$ occurs for reasons unrelated to Feldstein's theory, or 2) the change in expected inflation is even more difficult to measure than the level of expected inflation, and $\Delta\%P$ may be a more accurate measure than either the Livingston or the Fama series.

[7]In many recent macro models, economists have emphasized optimal forecasts (i.e., forecasts based on rational expectations), whereby rational economic agents utilize whatever information is available to obtain optimal forecasts. Many of the models that include optimal forecasts suggest that most, if not all, of the output effects of money growth and other policy variables result from errors in the forecasts of policy variables. Optimal forecasts are not directly observable, and no consensus has yet emerged about how to measure optimal forecast values. As a means of linking our analysis with the recent emphasis on forecast errors, we follow Karl Brunner, Alex Cukeirman, and Allan Meltzer (1980), who emphasize economic agents' inability to distinguish immediately between permanent and transitory shocks, and Stein, who emphasizes uncertainty and risk aversion as part of an "asymptotically rational expectations" hypothesis. In both of these models, expected money growth is at least partly determined by past money growth, which we proxy by $\%P$. Positive real balance growth implies faster than optimally forecast money growth and, given the other determinants of output growth, higher output growth; negative real balance growth implies slower than optimally forecast money growth and lower output growth.

A. *Linking Inflation with Future Output via the Real Balance Effect*

Fama (1981, 1982) and Fama and Michael Gibbons (1982) explained the inverse relation between inflation and future output as running from future (expected) output to current inflation. As described more fully below, Fama's inference that inflation responds inversely to next year's output is based on estimates of the following equation:[8]

$$(1)\quad \%P_t = \%P_t(\%B_t, \%Q_{t+1}, \%Q_t, \%Q_{t-1})$$
$$\%P_1 > 0; \quad \%P_{2,3,4} < 0,$$

where $\%P$ is the inflation rate, $\%B$ is the growth rate of the monetary base, and $\%Q$ is the growth rate of output. The purpose of this section is to show that the empirical estimates of Fama's inflation equation are more plausibly viewed as estimates of the following output (or unemployment) equation originally proposed by Stein:

$$(2)\quad \%Q_t = \%Q_t(\%B_{t-1}, \%P_{t-1}, U_{t-1})$$
$$\%Q_{1,3} > 0; \quad \%Q_2 < 0,$$

where U is the unemployment rate.

Parameter estimates of equation (1) are presented below. The estimates are obtained using annual data from 1954 to 1981. Similar results are obtained using other estimation periods such as 1954–76, 1956–76, and 1956–81. We measure $\%Q$ using annual average data for real *GNP*, $\%B$ using December data of the adjusted monetary base as published by the Federal Reserve Bank of St. Louis, and $\%P$ using December data of the personal consumption expenditures deflator. Except for the use of the consumption

[8]Fama also included $\%B_{t-1}$ in 1954–76 estimates of his inflation equation. We omit $\%B_{t-1}$ below because its coefficient was insignificant over the 1954–81 period used to estimate this and the other equations of this section. Estimates of the equations of this section over other periods, such as 1956–76 and 1956–81, are quite similar to those reported for 1954–81.

deflator rather than the *CPI*, our measures of $\%Q$, $\%B$, and $\%P$ are identical to Fama's. Estimates of equation (1), 1954–81, are as follows:

$$\underset{(4.70)}{\%P_t = 3.17} + \underset{(7.88)}{.73\%B_t} \underset{(-8.48)}{-.63\%Q_{t+1}}$$

$$\underset{(-3.70)}{-.26\%Q_t} + \underset{(.26)}{.01\%Q_{t-1}}$$

$$\bar{R}^2 = .91, \quad D\text{-}W = 2.02, \quad S.E. = .90, \quad \underset{(2.52)}{\rho = .48}$$

The coefficients of all of the variables in equation (1), except $\%Q_{t-1}$, are statistically significant. The first-order autocorrelation term is also significant. The significant first-order autocorrelation term reduces the standard error of the equation from .97 to .90 and suggests that a relevant variable has been omitted.

Of particular importance for Fama's purposes, the coefficient of $\%Q_{t+1}$ is $-.63$ with a t-statistic of -8.48. To explain how $\%P_t$ and $\%Q_{t+1}$ are related, Fama used a somewhat unconventional money demand model in conjunction with a "rational expectations version of the Fisherian quantity theory of money" (p. 549). In his money demand model, current money demand is a function not of current output, but of expected output which he proxied by future output ($\%Q_{t+1}$). (Fama viewed $\%Q_t$ and $\%Q_{t-1}$ as additional proxies of expected output.) In Fama's rational expectations quantity theory, output is taken to be determined completely independent of the monetary sector. This allows Fama to take future output (his proxy for expected output) as exogenous with respect to current money growth. Money supply is taken as exogenous with respect to prices, so that inflation is endogenous with respect to future output growth. Within this framework of money demand and the autonomous determination of output, Fama interpreted the strong negative relation between $\%P_t$ and $\%Q_{t+1}$ in equation (1) as the result of expected output (proxied by $\%Q_{t+1}$) affecting current money demand. Given nominal money growth, the output-induced movement in money demand must be accompanied by an inverse response of inflation in order to clear the equation of exchange.

Fama's approach is different from mainstream monetarist-neo-Keynesian analysis, both with respect to the demand for money and the quantity theory. With regard to money demand, Fama assumes as already mentioned that current money demand is a function of expected rather than current output. He suggests that this assumption follows directly from rational expectations and efficient market forecasts of future output. While market efficiency may suggest some incentive to base current money demand on expected (future) output, it does not relieve the public from the competing incentive to maintain money balances that are in equilibrium with respect to current output. Individuals will adjust money demand to expected output only if the costs of adjusting money balances coincident with output changes exceed the costs of holding balances out of equilibrium with current needs. Also, the positive response of money demand to higher expected output occurs in conjunction with another response already emphasized by Fama—namely, a positive response of stock prices to higher expected output. Unless the positive stock price response occurs completely independently of an increase in the demand for stocks, the positive responses of money demand and stock prices can occur only if an increase in expected output increases the demand for both money and common stocks. Such a portfolio adjustment would require a negative response in the demand for some third asset such as bonds. This is of course possible, but it would contrast with the response implied by most portfolio models which place bonds and most other assets intermediate between money at one extreme and real capital assets at the other. Finally, Fama includes an interest rate, measured at the beginning of the period, and finds its coefficient to be insignificant. Insofar as his inflation equation is the result of a process including money demand, it may be worth noting that almost all analyses of money

demand functions include a current interest rate as one of the arguments.[9]

Turning to the relation between inflation and output, Fama's assumption that output is determined independently of the monetary sector—or independently of money relative to prices—contrasts sharply with the disequilibrium dynamics of the real balance models discussed in the introduction to this section. In these models, monetary impulses affect the intermediate term path of output. Because of the differences between Fama's inflation-output relation and the causality posited in real balance models, we present parameter estimates of equation (2), which is a real balance/natural rate equation of output growth. Estimates are obtained using annual data, 1954–81. The level of unemployment U is measured by the married male unemployment rate which is more stationary than the total unemployment rate.

$$\%Q_t = \underset{(-1.79)}{-1.27} + \underset{(8.21)}{.81}\%B_{t-1} \underset{(-9.65)}{-.93}\%P_{t-1} + \underset{(6.94)}{1.33}U_{t-1}$$

$$\bar{R}^2 = .84, \quad D\text{-}W = 1.79, \quad S.E. = 1.02.$$

The coefficients on all of the variables of this equation are statistically significant and correct in sign. In contrast to equation (1), no autocorrelation correction is needed. Stein offered the following interpretation for equation (2). Changes in aggregate demand are related to lagged changes in real money balances. Equating expenditures and output, the growth rate of output ($\%Q_t$) is a function of the growth rate of lagged real balances ($\%(B_{t-1}/P_{t-1})$), which can be approximated by $\%B_{t-1} - \%P_{t-1}$. Reflecting the tendency of the economy to move toward a natural level of unemployment, real growth is also positively related to lagged unemployment, U_{t-1}.[10] By this interpretation, the statistically significant coefficients on lagged money growth and lagged inflation reveal a strong real balance effect. The statistically significant coefficient on lagged unemployment reflects the natural rate tendency of the economy.

Equations (1) and (2) are similar in that both contain base growth and inflation in one period and output growth one period later. This similarity, along with the unconventional nature of Fama's assumptions about money demand and the inflation-output relation, encouraged us to introduce the level of unemployment (contemporaneous with inflation) into equation (1). We removed $\%Q_{t-1}$ because it was insignificant, thereby obtaining

$$(3) \quad \%P_t = \%P_t(\%B_t, \%Q_{t+1}, \%Q_t, U_t).$$

Estimates of this equation, 1954–81, are as follows:

$$\%P_t = \underset{(1.62)}{1.24} + \underset{(14.99)}{.81}\%B_t \underset{(-11.69)}{-.77}\%Q_{t+1} \underset{(-2.91)}{-.20}\%Q_t + \underset{(3.19)}{.59}U_t$$

$$\bar{R}^2 = .93, \quad D\text{-}W = 1.83, \quad S.E. = .81.$$

The coefficient on unemployment is statistically significant. Of greater interest is the effect of unemployment on the total equation and virtually all of the individual parameters. With regard to the total equation, the standard error is reduced to .81 from .90, or more appropriately, from .97 before equation

[9]For a recent survey of alternative money demand specifications, all including at least one statistically significant interest rate coefficient, see John Judd and John Scadding (1982), especially Table 1, pp. 996–97.

[10]More precisely, real growth is positively related to lagged unemployment relative to its natural rate. We measured unemployment by a total unemployment rate linearly detrended and also by the married male unemployment rate, which had very little trend over the 1955–81 period. Because of the limited trend in married male unemployment, $U_t - U^*$ for married male unemployment is approximately equal to $U_t - U_{NAT}$, where U^* is the mean of U and U_{NAT} is the natural rate of unemployment. This allows us to interpret the coefficient on U in equation (2) not only as the coefficient on $U_t - U^*$ but also as the coefficient on $U_t - U_{NAT}$.

(1) was adjusted for first-order autocorrelation. The inclusion of unemployment removes the first-order serial correlation of the error term, suggesting that the omitted variable in equation (1) is either the level of unemployment or is highly correlated with the level of unemployment. The coefficients on both $\%B_t$ and $\%Q_{t+1}$ increase slightly, *while the t-statistics increase by* 90 *and* 40 *percent, respectively.* The jump in the t-statistics when unemployment is introduced suggests that the relation between these variables and inflation is more closely isolated by the inclusion of the level of unemployment. At the same time, the coefficient and t-statistic on current output falls, suggesting a tradeoff between current output growth and the current level of unemployment. *The opposite effects of unemployment on the coefficients and t-statistics of current output on the one hand, and future output on the other, suggests that current and future output bear very different structural relations to current inflation.*

Since the variables in equation (3) are almost identical to those in equation (2), the beneficial effects of U_t in equation (3) could almost have been anticipated once the similarity between Fama's inflation equation and Stein's real balance output equation was recognized. Whereas the importance of U_t in equation (3) follows directly from a real balance analysis, we can think of no reason within Fama's framework to explain the role of the level of unemployment. Particularly in view of the strong assumptions necessary to arrive at Fama's equation in the first place, we interpret equation (1) as a transformation of equation (2), the real balance output equation. By our interpretation, the estimates of equation (1) show a relation not of future output to current inflation via a money demand model of inflation but a relation of current inflation (and current money growth) to future output via a real balance model of output. *Fama has mistakenly transformed an output equation into an inflation equation.*

Stability tests of equations (1) and (2) provide additional support for our real balance/output as opposed to Fama's money demand/inflation interpretation. We estimated fourteen 15-year window regressions for both equations over the 1954–81 sample period. A "cusums" residuals test shows that equation (2) is stable; none of the residuals from the 15-year estimates fall outside a 95 percent confidence interval. In contrast, residuals from the same test of equation (1) fall outside the confidence intervals in 1962, 1965, and 1966. Moreover, all of the individual parameters of equation (2) are stable over the sample period, whereas two of the individual parameters of equation (1) show significant instability. In the early 15-year samples (1954–68 to 1962–76), the constant in equation (1) varies between 2.8 and 4.1, with a t-statistic always greater than 3.2. In the later samples (1964–78 through 1967–81), the constant in equation (1) varies between 1.1 and 2.1 with a t-statistic always below 1.6. The coefficient of monetary base growth in equation (1) moves from the .4–.5 area in the early 15-year samples (1954–68 to 1958–72) to .9–1.2 in the later samples (1964–78 to 1967–81). This instability of the inflation equation and two of its individual parameters contrasts with the stability of the output equation and *all* of its individual parameters.

B. *Linking Inflation to Real Stock Returns via the Real Balance Effect*

The relation of stock returns to *future* output growth and the relation of output growth to *lagged* base growth, inflation and unemployment in our real balance model of output suggest the following relation of stock returns to base growth, inflation and unemployment:

$$RS_t = RS_t(\%B_t, \%P_t, U_t)\, RS_{1,3} > 0;\; RS_2 < 0.$$

Estimates of this relation for 1954–76, 1956–76, and 1956–81 are presented below. The first interval is Fama's original period of estimation. The 1956–76 period was chosen because of evidence in Section I suggesting a change in the behavior of stock returns after 1955. The final interval was chosen to exclude 1954 and 1955 and to include the most recent years.

1954–76

$$RS_t = -.23 + 5.05\%B_t - 8.78\%P_t + 12.78U_t$$
$$\quad (-1.76) \quad (3.26) \quad (-5.50) \quad (4.26)$$

$$\bar{R}^2 = .76, \quad D\text{-}W = 1.92, \quad S.E. = .105, \quad \rho = .68 \; (3.67)$$

1956–76

$$RS_t = -.19 + 3.98\%B_t - 6.45\%P_t + 9.11U_t$$
$$\quad (-1.98) \quad (3.55) \quad (-5.09) \quad (3.96)$$

$$\bar{R}^2 = .69, \quad D\text{-}W = 1.70, \quad S.E. = .107$$

1956–81

$$RS_t = -.23 + 3.12\%B_t - 4.27\%P_t + 9.35U_t$$
$$\quad (-2.04) \quad (2.32) \quad (-3.39) \quad (3.42)$$

$$\bar{R}^2 = .43, \quad D\text{-}W = 1.86, \quad S.E. = .138$$

For all three time periods, the coefficients of all of the determinants are statistically significant and correct in sign. They suggest a negative relation between inflation and real stock returns, given base growth and unemployment. We interpret this partial relation as the result of inflation's effect on output via the real balance effect. The regressions of returns against the determinants of our output model also show a positive effect of base growth, given inflation and unemployment. Just as we interpreted the negative relation of real stock returns to inflation, given base growth and unemployment, as the result of inflation's negative effect on output via the real balance effect, we interpret the positive relation of real stock returns to base growth, given inflation and unemployment, as the result of base growth's positive effect on output via the real balance effect. This positive relation contrasts with Fama's finding of an inverse base growth-real stock return relation, which he dismissed as purely statistical and without economic significance. It contrasts even more sharply with Robert Geske and Richard Roll (1983), who suggest that the inverse relation between money growth and real stock returns is not a spurious statistical relation but instead reflects an inverse effect of real stock returns on money growth via the money supply process.

III. Summary and Conclusion

Most economists assume that, in long-run equilibrium, real variables such as real stock returns should be invariant with respect to nominal variables such as inflation. They also believe that, in efficient and forward-looking markets, real returns on equities should reflect expectations about real variables such as expected output growth. In seeming contradiction to market efficiency and the invariance between nominal and real variables, a negative simple correlation between real stock returns and contemporaneous inflation existed over the 1954–81 period, as well as over most of this century.

Fama (1981) argues that, in efficient and forward-looking markets, real returns should be based on expectations about real variables, such as future output. Any inverse inflation-real return correlation must reflect an inverse relation between inflation and future output growth, which Fama proceeds to explain as an effect of future output growth on current inflation via a money demand model. We have presented stronger support than Fama for his argument that, given future output growth, inflation exerts no independent effect on real stock returns. We also presented an alternative explanation of the inverse output-inflation relationship. According to our explanation, the inverse relationship between inflation and output runs from current inflation to future output via a real balance effect. Regardless of the direction of causality between inflation and output growth, our evidence and interpretation parallels Fama in suggesting that the stock market efficiently forecasts future output growth.

Our proposed relationship between inflation and real stock returns based on the real balance effect is fully consistent with long-run invariance between nominal and real variables as well as with market efficiency. This is because a structural relation between inflation and stock returns arising from the real

balance effect pertains only to periods of adjustment rather than long-run equilibrium. The real balance effect is a short-run disequilibrium phenomenon, operating only prior to the full adjustment of inflation to money supply growth.

REFERENCES

Benderly, Jason and Zwick, Burton, "Money, Unemployment and Inflation," *Review of Economics and Statistics*, February 1985, *67*, 139–43.

Bodie, Zvi, "Common Stocks as a Hedge Against Inflation," *Journal of Finance*, May 1976, *31*, 459–70.

Brunner, Karl, Cukierman, Alex and Meltzer, Allan H., "Stagflation, Persistent Unemployment and the Permanence of Economic Shocks," *Journal of Monetary Economics*, November 1980, *6*, 467–92.

Fama, Eugene F., "Stock Returns, Real Activity, Inflation, and Money," *American Economic Review*, September 1981, *71*, 545–64.

______, "Inflation, Output and Money," *Journal of Business*, April 1982, *55*, 201–31.

______ **and Gibbons, Michael R.,** "Inflation, Real Returns and Capital Investment," *Journal of Monetary Economics*, May 1982, *8*, 297–324.

______ **and Schwert, G. William,** "Asset Returns and Inflation," *Journal of Financial Economics*, November 1977, *5*, 115–46.

Feldstein, Martin, "Inflation and the Stock Market," *American Economic Review*, December 1980, *70*, 839–47.

Geske, Robert and Roll, Richard, "The Fiscal and Monetary Linkage between Stock Returns and Inflation," *Journal of Finance*, March 1983, *38*, 1–33.

Jaffe, Jeffrey and Mandelker, Gershon, "The 'Fisher Effect' for Risky Assets: An Empirical Investigation," *Journal of Finance*, May 1976, *31*, 447–58.

Judd, John and Scadding, John, "The Search for a Stable Money Demand Function: A Survey of the Post-1973 Literature," *Journal of Economic Literature*, September 1982, *20*, 993–1023.

Kessel, Reuben A., "Inflation-Caused Wealth Redistribution: A Test of a Hypothesis," *American Economic Review*, March 1956, *46*, 128–41.

Modigliani, Franco, "Monetary Policy and Consumption: Lingages Via Interest Rates and Wealth in the FMP Model," in *Consumer Spending and Monetary Policy: The Linkages*, Boston: Federal Reserve Bank of Boston, 1971, 9–85.

Nelson, Charles R., "Inflation and Rates of Return on Common Stock," *Journal of Finance*, May 1976, *31*, 471–83.

Shiller, Robert, "Theories of Aggregate Stock Price Movements," *Journal of Portfolio Management*, Winter 1984, *10*, 28–37.

Stein, Jerome L., *Monetarist, Keynesian and New Classical Economics*, New York: New York University Press, 1982.

The Cyclical Sensitivity of Wages

By ANTHONY O'BRIEN*

In recent articles in this *Review*, Jeffrey Sachs (1980) and Charles Schultze (1984)[1] have expressed differing views on the extent of the decline in the responsiveness of money wages to movements in aggregate demand in the post-World War II period. Each article presents a simple measure of wage responsiveness. While undoubtedly crude, these measures have a certain trout-in-the-milk appeal to them that contrasts with the unavoidable unpersuasiveness involved in using potentially more definitive, but actually more uncertain, specifications. Hence, resolving this disagreement would seem to be of some value.[2] This paper will demonstrate that, while Sachs' analysis contains certain errors, his main conclusion of significantly diminished postwar wage flexibility is correct. Schultze's contrary conclusions are shown to result from difficulties with the wage series he uses and from his selection of the time periods to be included in his analysis. Some additional data are also presented that indicate that severe wage rigidity in American manufacturing first appeared during the 1929–31 period. If true, this would indicate that the prewar/postwar dichotomy employed by Sachs, Schultze, and others is misleading, and would cast doubt on the argument often advanced[3] that multiyear wage contracts, or other postwar phenomena, account for the decline in wage responsiveness.

I

Tables 1 and 2 in Sachs' paper give a clear presentation of the declining cyclical responsiveness of wages in the postwar period. They are marred, however, by his use of a wage series that has been incorrectly prepared for the years 1890–1914. The result is that Sachs overstates the cyclical responsiveness of wages during these years.

The wage series employed in the wage inflation measures in Tables 1 and 2 (and used in calculating the dependent variable in the regressions presented later in the article) is, for the period 1890–1929, taken from Series B 72—Compensation per Hour, Manufacturing—in the Bureau of the Census *Long-Term Economic Growth, 1860–1965* (1966). The series was originally constructed by Albert Rees. Sachs describes his adjustments to the series as follows: "The Rees time-series is expressed in constant 1957 dollars. To get nominal wages, the wage series was multiplied by the *CPI*, Series E 135 in *Long-Term Economic Growth, 1860–1965*" (p. 89).

There are two problems with this. First—a minor point—there is no Series E 135 in the *Long-Term Economic Growth* volume. Sachs actually used Series E 135 in another Bureau of the Census publication, *Historical Statistics of the United States, Colonial Times to 1970*. More importantly, the source for Series B 72, Rees (1959, pp. 15–16) (see *Long-Term Economic Growth*, p. 150), gives the wage values in both real and nominal terms. For the years 1890–1914, the conversion from nominal to real values is made by Rees using

*Visiting Assistant Professor, Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, PA 15213. Information has been kindly supplied by Jeffrey Sachs and Charles Schultze concerning the methods and data used in their papers.

[1]Most of the evidence supporting Schultze's conclusions appears in his 1981 *Brooking Papers* article. All further references to Schultze's work are to this article.

[2]A definitive analysis that would resolve all disagreements concerning the cyclical responsiveness of wages would probably depend on agreement on the proper specification of a Phillips curve-like relation, or set of relations, that would allow for a clear measure of the impact of aggregate demand on wages. Anyone familiar with the debates on Phillips curve specifications can gauge the likelihood of this happening. In this particular context see, for example: Sachs' criticism of Michael Wachter's 1976 specification (pp. 82–83), Robert J. Gordon's criticism of Sachs (1980, pp. 244–45, 248, fn. 15), and Schultze's criticism of Gordon (1981, pp. 533, 589).

[3]See, for example: Gordon (1983, pp. 106–07), Schultze (1981, pp. 541–42) and Sachs (p. 88).

TABLE 1—COMPARISON OF THE SACHS AND REES WAGE SERIES

	Sachs	Rees
1890	.143	.144
1891	.143	.144
1892	.144	.145
1893	.151	.151
1894	.141	.139
1895	.137	.138
1896	.144	.144
1897	.140	.140
1898	.138	.137
1899	.146	.146
1900	.148	.151
1901	.153	.158
1902	.165	.165
1903	.172	.170
1904	.170	.169
1905	.173	.172
1906	.182	.184
1907	.188	.191
1908	.179	.184
1909	.182	.186
1910	.194	.198
1911	.196	.202
1912	.204	.207
1913	.220	.221
1914	.219	.220

Source: see text.

a cost-of-living index of his own construction.[4] Therefore, Sachs' method of multiplying Series B 72 by Series E 135 will not yield the correct values for 1890–1914 (25 of the 40 years in his pre-World War II sample). The series Sachs used and the Rees series he ought to have used are given in Table 1.[5]

Sachs bases his yearly cyclical dates on the National Bureau of Economic Research's quarterly dating, rather than on its yearly dating. Since the NBER considers its yearly dating to be more dependable than its monthly or quarterly dating, it might have been more standard to have used it.[6] Using the correct wage series, the NBER's yearly cyclical dating, exact percentage changes—rather than their logarithmic approximations—and a postwar series that, because it excludes overtime, is comparable to Rees' prewar average hourly earnings series would give the results in Table 2.[7] Sachs' general conclusion that there was a significant decline in the cyclical sensitivity of wages following World War II, while somewhat modified, would appear to still stand.[8]

Schultze has challenged Sachs' findings.[9] He argues that measured against movements in nominal *GNP*, there has been only a small decline in the cyclical sensitivity of wages in the postwar period. Schultze's conclusions rest on the data he presents in Table 6 (1981, p. 540), where he compares pre- and postwar flexibility coefficients for wages and finds

[4]Series E 135 is the BLS Consumer Price Index. The values for 1890–1912 are based on Rees. However, the technique used by the Bureau in splicing Rees' series to earlier and later wage series causes the values they report to differ from those given by Rees. Compare U.S. Department of Labor, Bureau of Labor Statistics, Bulletin 1790, p. 287, with Rees (1961, p. 4).

[5]The Rees series in Table 1 is from Rees (1961, p. 4) rather than from Rees (1959, pp. 15–16), since the former contains revised values for the years 1890–98. One other small point might be made here. In using Rees' total compensation series, rather than his average hourly earnings series, a difficulty arises in calculating the percentage change from 1928 to 1929. Rees only attempts to estimate wage supplements beginning in 1929; which is to say, before that year his average hourly earnings and total compensation series are identical. Therefore, the percentage change from 1928 to 1929 in the total compensation series is overstated. The total compensation series increases by 2.3 percent from 1928 to 1929, while the average earnings series increase by only 1.5 percent.

[6]See Geoffrey Moore (1961, p. 670) for the NBER's yearly dating. See p. 699 for a discussion of the greater dependability of the yearly dating over the quarterly and monthly dating. The years at issue are the cyclical peaks of 1892 and 1903, and the cyclical trough of 1896.

[7]The curiously small response of wages to cyclical activity during the 1890's that Sachs notes is actually made smaller yet in my Table 2. This apparent anomaly disappears if we break with the NBER and date the peaks and troughs for these years by the peaks and troughs in Rees' average hourly earnings series. Rees' series has peaks in 1893 and 1896, and troughs in 1895 and 1898 (1961, p. 4). The first two rows of Table 2 would then be amended as follows:

Year Before Peak to Peak (3)	Peak To Trough (4)	(4)−(3)
4.1	−4.3	−8.4
4.3	−2.4	−6.7

[8]The pre-World War II values for cols. 4–3 have a mean of −7.33 percent. The post-World War II values have a mean of −1.17 percent. This difference is significant at the 5 percent level for a one-sided test ($t=1.764$).

[9]Schultze (1981, especially pp. 541 and 571, and 1984, especially p. 160, fn. 1). See also the discussion between Sachs and Schultze in Schultze (1981, pp. 591–92).

TABLE 2—CYCLICAL SENSITIVITY OF MANUFACTURING WAGES

		Annual Percentage Changes Average Hourly Earnings, Manufacturing			
Year Before Peak to Peak (1)	Peak to Trough (2)	Year Before Peak to Peak (3)	Peak to Trough (4)	(4)–(3)	Percentage Difference in Cols. (4)–(3)[a]
1891–92	1892–94	.7	−2.1	−2.8	76.9
1894–95	1895–96	−.7	4.3	5.1	15.9
1898–99	1899–1900	6.6	3.4	−3.1	26.2
1902–03	1903–04	3.0	−.6	−3.6	14.3
1906–07	1907–08	3.8	−3.7	−7.5	10.7
1909–10	1910–11	6.5	2.0	−4.4	10.2
1912–13	1913–14	6.8	−.5	−7.2	1.4
1917–18	1918–19	32.0	14.4	−17.6	18.9
1919–20	1920–21	15.9	−11.8	−27.7	26.9
1922–23	1923–24	10.6	3.4	−7.2	7.5
1925–26	1926–27	.8	1.0	.2	–
1928–29	1929–32	1.5	−5.6	−7.1	–
1936–37	1937–38	11.8	−.5	−12.3	–
1943–44	1944–46	5.9	5.5	−.4	–
1947–48	1948–49	9.3	3.9	−5.4	–
1952–53	1953–54	5.7	3.0	−2.7	–
1956–57	1957–58	4.8	3.5	−1.2	–
1959–60	1960–61	3.3	2.7	−0.6	–
1968–69	1969–70	5.9	5.9	−.0	–
1972–73	1973–75	6.8	9.7	2.9	–

Source: Sachs, p. 80, Table 1. Data for 1890–1938 are average earnings per hour at work of wage earners in manufacturing. For 1890–1914: Rees (1961, p. 4). For 1917–38: Rees (1960, p. 3). Data for 1943–75 are average hourly earnings, excluding overtime, of production workers in manufacturing. For 1943–75: BLS Bulletin 1312–11, p. 930.

[a] From values in Sachs' Table 1.

only small declines. The flexibility coefficients consist of ratios of the difference between percentage changes per year in nominal wages (variously measured) in expansions and contractions to the difference between percentage changes per year in private nonfarm *GNP* in expansions and contractions. His calculations are flawed, however, by the inadequacy of all three of the prewar average hourly earnings measures he employs.

Schultze uses two series to represent prewar private nonfarm average hourly earnings: a series he constructed by dividing Stanley Lebergott's values (1964) for average annual earnings of full-time equivalent employees by "a series developed by Kendrick and the Bureau of Labor Statistics (BLS) on average hours worked" (1981, p. 539) and an index of wages appearing in W. S. Woytinsky.[10] One series is used to represent prewar average hourly earning in manufacturing. It consists of Rees' series on average hourly earnings per hour at work for 1900–19, and the BLS series on average earnings per hour paid for thereafter.[11] These series are all inferior to simply using Rees' series for the whole prewar period. The sources used by Lebergott are more appropriate to measures of annual earnings and, as Schultze notes, the technique used to convert Lebergott's values to hourly figures may result in an underestimate of the actual cyclical sensi-

[10] See Lebergott (p. 524) and Woytinsky (1953, p. 548).

[11] Schultze (1981, pp. 540–41); Rees (1959, pp. 15–16).

TABLE 3—FLEXIBILITY COEFFICIENTS

Period and Measure	Average Hourly Earnings (Manufacturing): Rees/BLS (1)	Rees (Including 1929–40) (2)	Rees (Excluding 1929–40) (3)
Prewar Peacetime Cycles			
Expansions	3.0	3.9	3.1
Contractions	0.4	−1.6	0.3
Change (Δw)	2.6	5.6	2.8
Flexibility Coefficients $\left(\frac{\Delta w}{\Delta y}\right)$	.25	.35	.26
	Adjusted Average Hourly Earnings		
	Adjusted Compensation	Private Nonfarm	Manufacturing
Postwar Peacetime Cycles			
Expansions	4.8	4.4	4.4
Contractions	4.0	3.7	3.3
Change (Δw)	0.8	0.8	1.1
Flexibility Coefficients $\left(\frac{\Delta w}{\Delta y}\right)$	0.11	0.10	0.16

Sources: Schultze (1981, Table 6, p. 540). Rees Series (1961, p. 4, and 1960, p. 3).

tivity of wages in the prewar period.[12] The Woytinsky series, although labeled by Schultze as "private nonfarm," includes, for 1900–13, earnings of federal and postal employees. After 1913 the Woytinsky series excludes federal and postal employees, but includes farm wage rates. Robert A. Gordon called the series "a hodgepodge," and referred to Rees' series as "more reliable" (1975, pp. 299–300).[13] Rees' series is also to be preferred to the prewar BLS series, since Rees corrects for the bias introduced into the BLS data by the underrepresentation of small firms in its sample.[14]

Substituting Rees' continuous series for the Rees/BLS hybrid that Schultze used appears to make a considerable difference in his results, as is indicated by comparing columns 1 and 2 of my Table 3. Schultze also appears to have made a computational error, since dividing his value for Δw (from the hybrid Rees/BLS series) with my value for Δy (constructed according to the procedure given by Schultze in his Appendix A) gives a result of .16. It seems a fair inference that the value for $\Delta w/\Delta y$ in prewar manufacturing that should have been reported by Schultze in his Table 6 is .16 rather than .25. However, in private correspondence, Schultze has provided information that clears up these discrepancies. A footnote was inadvertently dropped from Table 6 that would have stated that the data for the Great Depression years (1929–40) had been excluded from the analysis. Omitting these years results, by my

[12] Schultze (1981, p. 539, fn. 16). In discussing the construction of the manufacturing component of his index, Lebergott writes: "For 1900 to 1914, we do not use the recent estimates of Albert Rees. They mark a significant advance for the measurement of hourly earnings, which was his primary objective, but are less suitable for measuring annual earnings" (p. 184).

[13] See Woytinsky (p. 584, fn.a). I assume that Schultze used the series given in first column of Table 16, p. 584.

[14] See Rees (1960, pp. 5; 7–13; 15–19). The following discussion of the BLS series appears in U.S. Bureau of the Census, *Historical Statistics*...:

The estimates of average weekly earnings for 1909–1931, based primarily on census data, tend to be more accurate than those for average hourly earnings and average weekly hours. It is likely that the hourly earnings figures are overstated and the weekly hours understated because the BLS surveys of wages tended to sample large firms more heavily than small firms. [1975, p. 153]

calculation, in the values given in column 3 of my Table 3. The values for $\Delta w/\Delta y$ in columns 1 and 3 can be seen to be essentially the same. Accepting the value in column 3 as the appropriate one would result in the measure of wage flexibility having declined from the prewar period by from 38.46 to 61.54 percent, depending upon which postwar measure is used. However, it is difficult to justify excluding the observations from the Great Depression years from the analysis. Schultze's flexibility coefficients are designed to measure movements in nominal wages *per unit* of nominal *GNP* movement. Hence, it is not clear why a particular period ought to be excluded because of the severity of the nominal *GNP* movement that occurred during it. Moreover, Schultze's choice of years eliminates from his analysis three of the four cyclical contractions during the 1890–1940 period that were classified by Arthur Burns and Wesley Mitchell (1946) as "severe depressions."[15] Taking Alvin Hansen's (1964) classifications, Schultze has eliminated four of five severe depressions.[16] If, then, we use the value for $\Delta w/\Delta y$ for the prewar period given in column 2 as the appropriate one, the responsiveness of nominal wages to movements in nominal *GNP* declined by from 54.29 to 71.43 percent. If Schultze's flexibility coefficient is a meaningful measure of the sensitivity of wages to the business cycle, then its having more than halved from the prewar to the postwar period must surely indicate a substantial decline in the cyclical sensitivity of wages.[17,18]

II

In the discussions on the changing cyclical sensitivity of wages, not enough attention has been paid to the striking rigidity in nominal wages paid in manufacturing during the period from the cyclical peak of August 1929 to fall, 1931. There may well be a great deal of analytical insight to be gained from the study of this episode, occurring as it did against a background history of wage flexibility, and in an environment in which union strength and direct government intervention in labor markets were both quite limited. One difficulty in apprehending the distinctiveness of this period lies in the lack of monthly data for previous periods against which it can be compared.

It would be particularly useful to examine the behavior of wages following the January 1920 and August 1929 cyclical peaks. The comparison between 1920–21 and 1929–30 is likely to be illuminating because the two periods saw declines in economic activity of roughly equal magnitude. The decline in real *GNP* from 1920 to 1921 was 8.7 percent; from 1929 to 1930 it was 9.9 percent. There are available for making this comparison only two sets of monthly wage series, neither of which is wholly adequate. The average hourly earnings series published by the National Industrial Conference Board[19] only begin in June 1920—after the business cycle peak—and, somewhat mysteriously, omit the months January through June 1922, which include the cyclical trough in wages. The other series were compiled by Daniel Creamer (1950) from data on changes in wage rates contained in the payroll surveys of the Bureau of Labor Statistics. It seems clear that Creamer's

[15] Their discussion of severe depressions appears on pp. 455–64. Their calendar-year reference dates appear in Table 16, p. 78. Schultze omits 1892–94, 1920–21, and 1929–32; he includes 1907–08.

[16] Hansen (p. 24) counts 1937–38 as a severe depression.

[17] Schultze calls the decline in his wholesale price flexibility coefficient of 71.88 percent (p. 537) "substantial" (p. 571).

[18] Data problems in two other recent articles on this subject might be noted. Robert J. Gordon (1983) leaves some doubt as to the U.S. wage series he used, since he describes it (pp. 118–19) as having been constructed for the years 1890–1946 by multiplying Series B 70 by Series B 69 (in *Long-Term Economic Growth*). Since Series B 69 is a wholesale price index and Series B 70 is a consumer price index, this is presumably not what he did. Martin Bailey (1983), in his interesting discussion of the labor market in the 1930's, uses the manufacturing wage series in Rees (1959) rather than the revised series in Rees (1960).

[19] See Ada Beney (1937).

series understate the actual movements in wage rates. In responding to the BLS surveys, firms would sometimes not answer the question inquiring whether and by how much they had changed wage rates during the previous month. Creamer, following the BLS, interprets a failure by a firm to respond to the question as indicating that the firm had not made a wage rate change in the previous month. However, it is likely that some firms that failed to respond would in fact have changed wage rates. In addition, although the BLS surveys contain sufficient data to push series for eight industries back to October 1916, Creamer does not begin any of his series before January 1919. I have, therefore, derived average hourly earnings for eight manufacturing industries for the period October 1916 to May 1920 by using the BLS data to construct wage rate series that could be used to interpolate monthly values for yearly average hourly earnings data available in Paul Douglas (1930). These series were then spliced to the corresponding Conference Board series, with the January to July 1922 gaps filled by interpolations using the BLS data.[20]

Table 4 presents a comparison of movements in wages and wholesale prices in 1920–22 and 1929–35. The degree of money wage rigidity in 1929–31 appears striking. Using the NBER's cyclical dating, the average duration of contractions, measured from the cyclical peak to the following cyclical trough, during the period from 1890 to 1929 was 16.5 months. If the contraction beginning August 1929 had lasted 17 months, it would have ended in January 1931 with money wages in manufacturing having declined by only about 2 percent. During the same period, the Federal Reserve Board's seasonally adjusted index of total industrial production declined by 30.04 percent and the Board's seasonally adjusted index of manufacturing production declined by 32.35 percent.[21,22]

The comparisons in Table 4 between the magnitude of money wage declines in 1920–22 and in 1929–31 will be distorted by the fact that the 1920 downturn followed a period

[20]The industries used are automobiles, boots and shoes, hosiery and underwear, iron and steel, leather, paper, silk, and woolens. In 1919, these industries accounted for 18 percent of total manufacturing wage earners. The BLS survey results are from the *Monthly Labor Review*. I constructed the indexes using the method described in Creamer (Appendix A). Both the Douglas and Conference Board data appear to overstate the true average hourly earnings (see Rees, 1961, p. 37, and U.S. Bureau of the Census, 1975, p. 154) with the Douglas series being above the Conference Board series. An adjustment factor was used to make the Douglas values comparable to the Conference Board values. In deriving the final series, I used the method of interpolating time series by related series given in Milton Friedman (1962) and Friedman and Anna Schwartz (1970, pp. 321–28). (It should be noted that the series of average hourly earnings in the automobile industry appears to be somewhat dubious since it reaches a, probably spurious, peak in November 1918. It was retained in the analysis because movements in its values for 1919–20 seem consistent with wage movements in other industries, and because its elimination would have almost no effect on the results reported here.)

[21]Federal Reserve Board (1972). The nonseasonally adjusted index of manufacturing production declined by 35.17 percent during the same 17 months.

[22]The availability of monthly data on the operations of the U.S. Steel Corporation allow for an assessment of wage rigidity at the firm level. U.S. Steel made no general changes in wage rates during the Great Depression until a 10 percent cut was instituted on October 1, 1931. Shipments by U.S. Steel of rolled and finished steel products reached a peak of 1,702,000 short tons in May 1929. By October 1931, shipments had declined by 70 percent to only 520,000 short tons. Monthly net earnings by U.S. Steel also peaked in May 1929 at $25,605,000. Net earnings for October 1931 were only $1,690,000—a decline of 93 percent. Data on the common labor wage rate paid by U.S. Steel in its Pittsburgh district appear in U.S. Steel (1940, Vol. II, p. 182). From material in the iron and steel industry's trade journal, *Iron Age*, it is clear that movements in this wage rate represent almost exactly general movements in all wage rates at U.S. Steel. Which is to say, announcements in the *Iron Age* of wage rate changes at U.S. Steel match in magnitude and date changes in the common labor wage rate. See, for example, *Iron Age*, September 24, 1931, p. 836. Shipments are from U.S. Department of Commerce, *Survey of Current Business*, April 1940, p. 18. Net earnings are defined as "earnings after deduction of interest on bonds of subsidiary companies, but before charges and allowances for depletion, depreciation, amortization, and obsolescence" (*Survey of Current Business, 1936 Supplement*, p. 119). The earnings data are from *1931 Supplement*, p. 81, and *1932 Supplement*, p. 213.

TABLE 4—MOVEMENTS IN AVERAGE HOURLY EARNINGS IN MANUFACTURING AND IN WHOLESALE PRICES, 1920–22 AND 1929–31

	Wage or Price Peak	6 Months After Peak	One Year After Peak	Wage or Price Trough	
1920–22					
Average Hourly Earnings	(11/20)	(5/21)	(11/21)	(3/22)	
8 Industries[a]	.613	.512	.463	.436	
Percent Change	–	–16.48	–24.47	–28.87	
Average Hourly Earnings	(10/20)	(4/21)	(10/21)		
25 Industries[b]	.611	.545	.492	[c]	
Percent Change	–	–10.80	–19.48	–	
Wholesale	(5/20)	(11/20)	(5/21)	(1/22)	
Price Index[d]	167.2	133.4	96.2	91.4	
Percent Change	–	–20.22	–42.46	–45.33	
1929–31	Wage or Price Peak	One Year After Peak	January 1931	Two Years After Peak	Wage or Price Trough
Average Hourly Earnings	(8/29)	(8/30)	(1/31)	(8/31)	(3/33)
8 Industries[a]	.572	.568	.560	.549	.423
Percent Change	–	–0.70	–2.10	–4.02	–26.05
25 Industries[b]	.590	.591	.578	.568	.459
Percent Change	–	+0.17	–2.03	–3.73	–28.54
Wholesale	(8/29)	(8/30)	(1/31)	(8/31)	(2/33)
Price Index	97.7	84.0	78.2	72.1	59.8
Percent Change	–	–14.02	–19.96	–26.20	–38.79

[a] These are the eight industry series described in the text. They were aggregated using census employment weights for 1919.
[b] Beney (pp. 44–47).
[c] There are gaps in the Conference Board series for the months January to June 1922. This period appears to include the cyclical trough in wages.
[d] BLS Wholesale Price Index (1926 = 100).

of significant price inflation, while the 1929 downturn followed a period of relative price stability. However, a comparison of the *timing* of wage cuts relative to price cuts need not be affected. It seems reasonable to characterize an increase in the lag of wage cuts behind price cuts as an increase in wage rigidity. If price reductions reflect firm recognition of weakness in the product market, then a significant increase in the lag of wage cuts behind price cuts indicates a change in firm attitudes towards the role wage cuts should play in meeting a decline in demand. For those eight industries for which full wage data are available, the mean lag of the peak in a 3-month moving average of average hourly earnings behind the peak in a 3-month moving average of the corresponding industry wholesale price index was 2.25 months in 1920 and 13.69 months in 1929. This difference is statistically significant at the 1 percent level ($t = 4.74$). There are eight additional industries for which both Conference Board monthly wage data and monthly wholesale price data are available.[23]

[23] The difficulty in using this latter group of eight industries is that, since no observations exist for the months January to May 1920, the occurrence of a wage peak during these months will not be detected. Hence, for these industries the extent of the lag of wage peaks behind price peaks in 1920 is likely to be overstated, and, therefore, any change in cyclical behavior from 1920 to 1929 will be understated.

Taking all sixteen industries, the mean lag of wage peaks behind price peaks was 3.03 months in 1920 and 10.84 months in 1929. This difference is also significant at the 1 percent level ($t = 4.32$).

My suspicion is that the wage rigidity of 1929–31 represents a distinct break with previous wage response. The relative promptness of wage cutting in 1920—but not, it is worth repeating, the magnitude of the cuts—is likely to have been typical of earlier significant downturns. It has been argued, however, that the wage flexibility of 1920–22 was atypical and can be explained by its having been preceded by a wartime inflation. For instance, in Schultze's view:[24]

> The rapid and unusual fall in inflation in the recessions immediately following the two [world] wars is, I think, instructive. It is...consistent with a rational expectations view of the world. Precisely because both were, and more importantly, were seen to be, unique events, neither the inflation they generated nor the monetary policies that accompanied them were extrapolated into the future. The first sign of major weakness in economic activity led to a collapse of price expectations.... And even though labor markets eased somewhat when the war ended, they remained relatively tight into the immediate postwar boom. But the very high inflation rates did not get built into the "norm" that underlay implicit wage contracts. Sharp reductions in the rate of wage increases in the recessions of 1920 and 1949 did not violate those implicit wage contracts. [1981, p. 563]

The point is arguable, but is probably not capable of being resolved in the absence of monthly data for previous periods. However, it is worth noting that the circumstances of wage cutting in manufacturing in 1920 do not quite match the account generally given. For the eight industries for which wage series have been constructed as recounted above, wholesale prices peaked on average in May 1920, and wages peaked in August.[25] The wholesale price index, which then, as now, was made available by the BLS on a monthly basis, also peaked in May. By August it had declined by only 3.47 percent. Hence, wage cuts had begun *before* an accurate gauge of the extent of the severity of the contraction and of the accompanying deflation was possible. The economy had passed through the mild recession of late 1918–early 1919 without serious deflation,[26] and there is no obvious reason why the severe deflation that in fact occurred in 1920–22 need have been anticipated.

If, in fact, the decline in the cyclical responsiveness of wages began in 1929, then this poses an interesting conundrum for macroeconomic theorists. Those who are convinced that the large unexploited opportunities for gain that seemingly result from wage rigidity disappear provided that the optimizing behavior of firms and workers is correctly specified[27] will, I suspect, have great difficulty in identifying the changes in choice-theoretical fundamentals during the 1920's that resulted in the onset of nominal wage rigidity. Those who prefer to impose their structure well above the choice-theoretical bedrock will have a similar difficulty in identifying any institutional changes sufficient to have brought about the rigidity.

[24] Robert J. Gordon (1983, p. 118) makes a similar argument.

[25] By this I mean that the mean lag of peaks in 3-month moving averages of wholesale prices for these industries behind the cyclical peak of January 1920 was 3.63 months. As already noted, the mean lag of wage peaks behind price peaks was 2.25 months.

[26] The eight-industry average hourly earnings series declined by 4.93 percent from November 1918 to May 1919. The WPI declined by 5.60 percent from September 1918 to February 1919. This decline in wages was not previously observable; Rees' yearly average hourly earnings measure increases 14.39 percent from 1918 to 1919. (Taking monthly averages, my eight-industry measure increases by 18.95 percent from 1918 to 1919.)

[27] I am thinking here of the implicit contracts literature of Costas Azariadis (1975) and others. Since the equilibria characterized in these papers are generally not Pareto efficient (in the absence of the invoking of some form of a nonsubstitution principle; for example, Douglas Gale 1983, ch. 5, and Azariadas, pp. 1198–200), it is not quite correct to say that the problem of unexploited opportunities for gain is entirely taken care of.

REFERENCES

Azariadis, Costas, "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, December 1975, *83*, 1183–202.

Bailey, Martin N., "The Labor Market in the 1930s," in J. Tobin, ed., *Macroeconomics, Prices, and Quantities: Essays in Honor of Arthur M. Okun*, Washington: The Brookings Institution, 1983.

Beney, Ada M., *Wages, Hours, and Employment in the United States, 1914–1936*, New York: National Industrial Conference Board, 1937.

Burns, Arthur F. and Mitchell, Wesley C., *Measuring Business Cycles*, NBER *Studies in Business Cycles*, No. 2, New York: Columbia University Press, 1946.

Creamer, Daniel, *Behavior of Wage Rates During Business Cycles*, NBER Occasional Paper 34, New York, 1950.

Douglas, Paul H., *Real Wages in the United States, 1890–1926*, Boston: Houghton Mifflin, 1930.

Friedman, Milton, "The Interpolation of Time Series by Related Series," *American Statistical Association Journal*, December 1962, *57*, 729–57.

______ **and Schwartz, Anna J.,** *Monetary Statistics of the United States: Estimates, Sources, and Methods*, NBER *Studies in Business Cycles*, No. 20, New York: Columbia University Press, 1970.

Gale, Douglas, *Money: In Disequilibrium*, Cambridge: Cambridge University Press, 1983.

Gordon, Robert A., "Wages, Prices, and Unemployment, 1900–1970," *Industrial Relations*, October 1975, *14*, 273–301.

Gordon, Robert J., "A Consistent Characterization of a Near-Century of Price Behavior," *American Economic Review Proceedings*, May 1980, *70*, 243–49.

______, "A Century of Evidence of Wage and Price Stickiness in the United States, the United Kingdom, and Japan," in J. Tobin, ed., *Macroeconomics, Prices, and Quantities: Essays in Honor of Arthur M. Okun*, Washington: The Brookings Institution, 1983.

Hansen, Alvin H., *Business Cycles and National Income*, expanded ed., New York: W. W. Norton, 1964.

Lebergott, Stanley, *Manpower in Economic Growth: The American Record Since 1800*, New York: McGraw-Hill, 1964.

Moore, Geoffrey H., *Business Cycle Indicators*, Vol. I, Princeton: Princeton University Press, 1961.

Rees, Albert, "Patterns of Wages, Prices and Productivity," in *Wages, Prices, Profits, and Productivity*, New York: The American Assembly, 1959.

______, *New Measures of Wage-Earner Compensation in Manufacturing, 1914–57*, NBER Occasional Paper 75, New York, 1960.

______, *Real Wages in Manufacturing, 1890–1914*, Princeton: Princeton University Press, 1961.

Sachs, Jeffrey, "The Changing Cyclical Behavior of Wages and Prices: 1890–1976," *American Economic Review*, March 1980, *70*, 78–90.

Schultze, Charles L., "Some Macro Foundations for Micro Theory," *Brookings Papers on Economic Activity*, 2:1981, 521–92.

______, "Cross-Country and Cross-Temporal Differences in Inflation Responsiveness," *American Economic Review Proceedings*, May 1984, *74*, 160–65.

Wachter, Michael, "The Changing Cyclical Responsiveness of Wage Inflation," *Brookings Papers on Economic Activity*, 1:1976, 115–68.

Woytinsky, W. S., *Employment and Wages in the United States*, New York: Twentieth Century Fund, 1953.

Board of Governors of the Federal Reserve System, *Industrial Production, 1971 Revision*, Washington, 1972.

U.S. Department of Commerce, Bureau of the Census, *Long-Term Economic Growth, 1860–1965*, Washington: USGPO, 1966.

______, *Historical Statistics of the United States, Colonial Times to 1970*, Washington: USGPO, 1975.

U.S. Department of Labor, Bureau of Labor Statistics, *Employment and Earnings: United States, 1909–71*, Bulletin 1312–8, Washington: USGPO, 1972.

______, *Handbook of Labor Statistics, 1978*, Bulletin 2000, Washington: USGPO, 1979.

U.S. Steel Corporation, *T.N.E.C. Papers*, 3 vols., New York: U.S. Steel Corporation, 1940.

The Implications of Spillover for the Design of Monetary Policy: An Empirical Analysis of Income and Price Determination in Nonclearing Markets

By J. DAVID FERGUSON AND WILLIAM R. HART*

Macroeconomics is currently dominated by two competing frameworks of analysis through which the adherents of each attempt to explain the paths of aggregate economic variables and address the design of policy. One, the market-clearing or equilibrium approach to business cycles, is most prominently associated with the work of Robert Lucas (1972), Thomas Sargent (1976), Sargent and Neil Wallace (1976), and Robert Barro (1976). The other employs the non-market-clearing framework made fashionable by the work of Stanley Fischer (1977), Edmund Phelps and John Taylor (1977), and Taylor (1979).

Despite the apparent theoretical superiority of the market-clearing approach (i.e., the realization of all perceived gains from trade), the non-market-clearing approach continues to dominate analysis of macroeconomic fluctuations. This is explained partly by the failure of equilibrium macroeconomic models to fare well empirically and partly by the successful incorporation of the notion of rational expectations into non-market-clearing models.

The current spirit of the non-market-clearing approach emphasizes forward-looking (rational) contractual wage setting in which wages are set to clear labor markets. Short-run rigidity of the wage contract, however, prevents instantaneous clearing of the labor market in the face of economic disturbances, and as such, becomes a source of short-run deviations in aggregate variables from their natural rates.

One aspect of the non-market-clearing approach emphasized early in the literature by Don Patinkin (1952) is the notion of spillover—a situation in which buyers facing markets that clear only gradually are unable to purchase all they intend of a good at a given price, and therefore, redirect part of their unspent income to another market. The micro foundations for spillover in a model of nonclearing markets were provided in a seminal paper by Herschel Grossman (1969) in which Patinkin's concept of spillover was synthesized with the Takashi Negishi-Frank Hahn (1962) nontatonnement transaction process and Robert Clower's (1965) dual-decision hypothesis. And yet, despite these contributions, the concept of spillover has been largely ignored in the majority of current genre non-market-clearing models.

In two companion pieces (1980, 1983), we formally incorporated Grossman's (1969, 1971) work into a generalized macro model and used dynamic simulations to derive the implications of market spillover for income and interest rate determination in a short-run environment of gradual adjustment of demand and expectations and nonclearing markets. Our results suggested that spillover provides an additional potential explanation for short-run quantity adjustments and, as such, is important for the short-run design of monetary policy. However, two considerations serve to temper our earlier results: 1) the models employed were fixed price; and 2) the models were not estimated. Therefore, the question still remained whether spillover mattered empirically.

This paper answers this question by specifying and estimating a generalized macroeconomic model that not only incorporates the concept of spillover into a framework of nonclearing financial and real markets, but also includes a supply side to allow for flexi-

*Associate Professors, Miami University, Oxford, OH 45056. We thank James Dunlevy, William Hutchinson, and Nicholas Noble for their generous help on this paper. And special thanks are due Bill McKinstry for his continuous support and encouragement. We alone are responsible for any errors.

ble, slowly adjusting prices, rational expectations, and natural levels of the real variables. This model, therefore, allows for the estimation of both the level and direction of spillover as well as the rates of adjustment of demands and markets toward their equilibrium values.

Our results are important for three reasons. First, the empirical analysis is carried out within a theoretical context which builds upon current nonclearing rational expectations models (and which Grossman, 1983, argues is necessary for convincing analysis). Second, they represent (to our knowledge) the first successful estimation of spillover coefficients, and third, they show that spillover from the bond to the commodity market is indeed empirically significant, and thus is important for the short-run design of macroeconomic policy.

The outline of this paper is as follows. In Section I, the generalized model is presented and discussed. The estimation procedure and empirical results are reported in Section II, and we finish with a discussion of the significance of our estimates of the dynamic parameters and spillover coefficients.

I. The Model

The macroeconomic model specified below is a continuous dynamic system of equations designed for short-run analysis which assumes three goods: aggregate output, bonds, and money. These goods are traded in two markets that adjust only gradually to determine aggregate output, prices, and interest rates.

$$(1) \qquad y^* = s + y_r r^e + y_y Y^e + Y_\pi \pi e$$

$$(2) \qquad m^* = h + m_r(r^e + \pi^e) + m_y Y^e$$

$$(3) \qquad Dy = 1/j(y^* - y)$$

$$(4) \qquad n = 1/k(m^* - M)$$

$$(5) \qquad DY = 1/\lambda\{\min(\bar{y}', Y_N - Y)\}$$

$$(6) \qquad \pi = 1/w(\bar{y}_N') + \pi^e$$

$$(7) \qquad Dr = -1/\mu(\bar{b}')$$

$$(8) \qquad \bar{b}' = \bar{b} + \alpha \bar{y}$$

$$(9) \qquad \bar{y}' = \bar{y} + \beta \bar{b}$$

$$(10) \qquad \bar{n} + \bar{b} + \bar{y} = 0$$

where the parameters of the model are expected to take on the following signs:

$$m_r, y_r, y_\pi < 0$$

$$y_y, m_y, j, k, \lambda, w, \mu > 0$$

$$0 \le \alpha, \quad \beta \le 1$$

y^* = the desired or ideal private real demand for national product, y = actual current private real demand for national product, m^* = the ideal real money balances, n = the actual demand for additions to real money balances, b = the demand for additions to bond holdings, r = the nominal interest rate, r^e = the expected real rate of interest, Y = real national product, Y^e = real expected national product, Y_N = the real level of output consistent with the natural rate of unemployment, π = the rate of change of the absolute price level, π^e = the expected rate of change of prices, g = government demand for national product plus net exports, M = the real money supply, B = real flow supply of bonds, and D = the time derivative operator.

The model is specified to allow analysis of the processes of adjustment of 1) the discrepancies between actual and ideal demands for output and money (equations (1)–(4), and 2) prices and quantities in the real output and financial markets (equations (5)–(7)). It is assumed that the interaction of market transactions and the demand processes which do not clear instantaneously are best described here by the fixed-price equilibrium process developed by Grossman (1971), in which both effective and realized transactions are determined simultaneously (equations (8)–(9)).[1]

[1]Since this paper is designed to study the processes of demand and market adjustment that produce spillover, we consider only short-run behavior, even though this system of differential equations clearly has longer-run implications. See John Pippenger (1982) for development of the long-run analysis of such a system.

Notional and effective demands for national product, bonds, and money balances are represented by y, b, n, and y', b', n', respectively. Notional demands are the quantities of goods individuals desire to purchase or sell at current price levels. They are derived by Clower subject to the individual's budget constraint from an implicit utility-maximization process. Grossman (1971) defines effective demands as not equal to notional demands when markets are out of equilibrium, and are slow to clear in response to excess notional demand for a good. In the face of nonclearing markets, both individuals and firms must adjust their demands in the light of what can actually be transacted, and then decide what portion of their excess demand will be redirected into other markets. This recursive process will simultaneously determine effective demands in each market. Therefore, effective demand will include notional demand plus any spillover, or excess notional demand for other goods which is redirected into the market. Excess notional demands for national product $(y+g-Y)$, bonds $(b-B)$, and additions to money balances $(Dm-DM)$ are indicated by $\bar{y}$, $\bar{b}$, and $\bar{n}$, respectively; and the effective excess demands are designated by $\bar{y}'$, $\bar{b}'$, and $\bar{n}'$. Actual realized transactions in our model of "sticky" prices are then, as shown by Grossman (1974), equal to the minimum of effective demand and supply.

The first four equations of the model decribe a dynamic theory of demand for nominal output and additions to money balances. Equations (1) and (2) specify how individuals and firms select their ideal notional demands for national product (y^*) and ideal money balances (m^*), given their budget constraint (Y^e) and real expected interest rates (r^e), and the expected rate of inflation (π^e).[2] In equation (2), m^* is assumed to be a function of the sum of r^e and π^e, the nominal rate of interest.[3] In equations (3) and (4), costs of renegotiating existing contracts $(j>0)$ and obtaining money $(k>0)$ produce only gradual adjustment of this instant's actual notional demands, y and n, toward the ideal demands.

Both private consumption and investment demands are included in y, and for simplicity, federal taxes are subsumed in national income and the government budget constraint is not explicitly considered. Equation (4) specifies the adjustment of the notional flow demand for additions to money balances in the face of differences between the ideal and actual money stocks.[4]

Equations (5), (6), and (7) represent the market-clearing process for output, prices, and interest rates. It is assumed that actual national output adjusts gradually in response to the minimum of (a) excess effective demand, or (b) the excess of the natural level of income over actual current output (equation (5)), with the parameter λ indicating the speed of adjustment. A specification similar to equation (5) is suggested by Barro and Grossman (1971) as a way to allow for the complication of inventory adjustment. The use of Y_N in equations (5) and (6) incorporates a supply side into the model in order to make scarcity explicit.[5] Equation (6) expresses the gradual adjustment of prices $(w>0)$ in response to excess of effective demand over the natural level of output plus price expectations.[6] Equation (7) states that interest rates change gradually $(\mu>0)$ in re-

[2]The real expected rate of interest was calculated from the complete version of the Irving Fisher (1896, 1930) equation: $r=r^e+\pi^e+r^e\cdot\pi^e$. Thus, $r^e=r-\pi^e/1+\pi^e$.

[3]In theory, π^e might also be expected to enter the m^* function as a separate variable, but empirical tests showed its contribution to the explanation of m^* to be insignificant and highly correlated with $(r^e+\pi^e)$. The form $r^e+\pi^e$ is used to represent nominal interest rates because it properly reflects the impact of the cross product $(r^e\cdot\pi^e)$ in the Fisher equation.

[4]This equation can be thought to combine the relationship (a) between the ideal stock and the ideal rate of change in the stock of money, and (b) a relationship analogous to equation (3), between an actual and ideal flow demand. Because y^* is dimensioned as a flow, and the first of these relationships for national product is subsumed in equation (1).

[5]Robert Hall (1983) presents an excellent intuitive discussion of the natural rate hypothesis.

[6]This equation is consistent with the Lucas and Leonard Rapping (1970) presentation of the Phillips curve, but emphasis is upon the inverse relationship of price determination rather than employment or output determination.

sponse to disequilibrium in the bond market. The excess of flow demand for bonds over supply is expected to produce falling market interest rates.

Finally, equations (8) and (9) represent simplified versions of the relationship between aggregate effective demand and notional demand as derived from its micro foundations by Grossman (1971). The excess effective demand for bonds, equation (8), is equal to the sum of the excess notional demand for bonds $\bar{b}$, plus $\alpha\bar{y}$, the spillover of unfulfilled notional demand from the goods market. The parameter α is the spillover coefficient. Likewise, equation (9) represents the equality between the excess effective demand for output of goods and the sum of excess notional demand $\bar{y}$, plus the spillover of unfulfilled excess notional demand for bonds, $\beta\bar{b}$, where β is the coefficient of spillover. Both α and β are assumed constant, and in the range of 0 (no spillover) to 1 (complete spillover of excess notional demand).[7]

II. Empirical Results

The equations of the model were reduced to a form suitable for estimation in which only the observable variables Y, M, r, π, Y_N, g, Y^e, r^e, π^e, and t were present. Since M and g were assumed exogenous policy variables, Y_N, r^e, and π^e exogenous data to the model, and Y^e was estimated separately from the model, the system was reduced through substitution to three simultaneous differential equations describing the generation of Y, Dr, and π. However, the conditional nature of equation (5) obviously complicates the reduced-form equation describing the generation of Y. The choice of the appropriate equation depends upon the value of y, a nonobservable variable. Thus, to establish the appropriate equation at any given time, a test was needed that was dependent upon an observable variable to determine which form of the equation was appropriate to represent the output market. The relationship between π and π^e provides such a test.[8]

When $\pi \leq \pi^e$, the appropriate form of equation (5) is

$$DY = 1/\lambda[(y+g)-Y]. \tag{5'}$$

If $\pi > \pi^e$, the appropriate form of equation (5) is

$$DY = 1/\lambda[Y_N - Y]. \tag{5''}$$

Thus, when $\pi \leq \pi^e$, the reduced-form of the model is as follows:[9]

$$\begin{aligned} Y = {} & a_1D^2Y + a_2DY + a_3Dg + a_4g \\ & + a_5DM + a_6M + a_7Dr^e + a_8r^e \\ & + a_9DY^e + a_{10}Y^e + a_{11}D\pi^e + a_{12}\pi^e + a_{13} \end{aligned} \tag{11}$$

or if $\pi > \pi^e$,

$$Y = Y_N - \lambda DY. \tag{11'}$$

$$\begin{aligned} Dr = {} & b_1D^2r + b_2D^2M + b_3DM \\ & + b_4M + b_5DY + b_6Y + b_7Dg \\ & + b_8g + b_9Dr^e + b_{10}r^e + b_{11}DY^e \\ & + b_{12}Y^e + b_{13}D\pi^e + b_{14}\pi^e + b_{15} \end{aligned} \tag{12}$$

$$\begin{aligned} \pi = {} & c_1D\pi + c_2DY + c_3Y + c_4DY_N \\ & + c_5Y_N + c_6D^2M + c_7DM + c_8M \\ & + c_9Dg + c_{10}g + c_{11}Dr^e + c_{12}r^e \\ & + c_{13}DY^e + c_{14}Y^e + c_{15}D\pi^e + c_{16}\pi^e + c_{17} \end{aligned} \tag{13}$$

[7]Our earlier articles (1980, 1983) derive the theoretical implications of the polar values of these spillover coefficients within a generalized model with a fixed price level.

[8]If $y' + g < Y_N$, the change in output is determined by the size of excess effective demand (y'), but if $y' + g > Y_N$, the change in output is determined by $Y_N - Y$. It can be seen from equation (6) that $\pi - \pi^e = 1/w(y' + g - Y_N)$. Therefore, if $\pi > \pi^e$, then $y' + g > Y_N$; and if $\pi \leq \pi^e$, then $y' + g \leq Y_N$. Thus, the relationship between π and π^e provides a test of the appropriate form of equation (5) based upon observable criteria.

[9]An appendix is available upon request which presents the complicated substitutions necessary for the derivation of the reduced form of the model.

where

$a_1 = -j\lambda(1-\beta)$, $a_2 = -[j+\lambda/(1-\beta)]$

$a_3 = j$, $a_4 = 1$, $a_5 = j\beta/k(1-\beta)$, $a_6 = \beta/k(1-\beta)$

$a_7 = -jm_r\beta/k(1-\beta)$, $a_8 = y_r - \beta m_r/k(1-\beta)$

$a_9 = -jm_y\beta/k(1-\beta)$, $a_{10} = y_y - [\beta m_y/k(1-\beta)]$

$a_{11} = -jm_r\beta/k(1-\beta)$, $a_{12} = y_\pi - \beta m_r/k(1-\beta)$

$a_{13} = s - \beta h/k(1-\beta)$

$b_1 = -j$, $b_2 = j/\mu$, $b_3 = 1/\mu(1-j/k)$

$b_4 = -1/\mu k$, $b_5 = j/\mu$, $b_6 = 1/\mu$

$b_7 = -j(1-\alpha)/\mu$, $b_8 = -(1-\alpha)/\mu$

$b_9 = jm_r/\mu k$, $b_{10} = (1/\mu)[m_r/k - y_r(1-\alpha)]$

$b_{11} = jm_y/\mu k$, $b_{12} = (1/\mu)[m_y/k - y_y(1-\alpha)]$

$b_{13} = jm_r/\mu k$, $b_{14} = (1/\mu)[m_r/k - y_\pi(1-\alpha)]$

$b_{15} = h/\mu k - (1-\alpha)s/j\mu$

$c_1 = -j$, $c_2 = j/w$, $c_3 = 1/w$, $c_4 = -j/w$

$c_5 = -1/w$, $c_6 = j\beta/w$, $c_7 = (\beta/w)(j/k+1)$

$c_8 = \beta/wk$, $c_9 = -j\beta(1-\beta)/wk$

$c_{10} = -\beta(1-\beta)/wk$, $c_{11} = -j\beta m_r/wk$

$c_{12} = 1/w[(1-\beta)y_r - \beta m_r/k]$, $c_{13} = -j\beta m_y/wk$

$c_{14} = (1/w)[y_y(1-\beta) - m_y\beta/k]$

$c_{15} = j[1-\beta m_r/wk]$

$c_{16} = 1 + 1/w[(1-\beta)y_\pi - \beta m_r/k]$

$c_{17} = 1/w[(1-\beta)s - \beta h/k]$

The Statistical Analysis System (SAS) Institute (1980) *SYSNLIN*, an iterated three-stage least squares program designed for nonlinear simultaneous systems, was used to estimate the parameters of the model. Because the model is in continuous form, this presented a problem in the use of data reported in standard form for time aggregates since the quantities are measured as average values over a finite time period. Therefore, this paper uses a nonrecursive approximation of the continuous model as suggested by A. R. Bergstrom (1966).[10]

The model was estimated using data for the U.S. economy chosen to include the period from the first quarter of 1959 through the fourth quarter of 1980. The variables of the model were defined as follows: r = twelve-month U.S. Treasury note rate; Y = *GNP*, 1972 U.S. dollars; g = federal, state and local government purchases, 1972 dollars; M = currency plus transaction deposits ($M1$), 1972 dollars; Y_N = natural *GNP*, 1972 dollars; r^e = expected real twelve-month Treasury bill rate; Y^e = permanent income, 1972 dollars,[11] and π^e = expected rate of price change, *CPI*.

Data for r and M were taken from the *Federal Reserve Bulletin*, Y and g from the

[10] The continuous form was chosen for this study for two basic reasons: it presents a better picture of aggregate behavior; and the expected advantages of a discrete model, when only periodic data are available, are not as promising as they would appear at first glance. When aggregation is made over many decision units and types of goods, a continuous model seems to present a better approximation of actual behavior. Both individuals on the demand side of the market and suppliers of goods would not be expected to be homogenous in their behavior. Reaction times would differ, and after the initial delayed response until the first individual reacted to the market, the overlapping of many individuals and goods in the market would obscure the discrete behavior of each unit. Thus, a continuous system would seem to give an appropriate representation of aggregate response. Yair Mundlak (1961) has shown that, unless the average inaction period in a discrete model equals the observation period of the data, incorrect estimates of the adjustment coefficients will result. This will happen even if the system is truly discrete. Since all data used in this work are reported quarterly, and the reaction time is surely much shorter than this, there is no advantage to using a discrete model.

[11] Permanent income data was constructed using the procedure suggested by Michael Darby (1974) as specifically applied by Anthony Santomera and John Seater (1981). The following equation was used for this process: $Y^e = aY_t + (1-a)(1+b)Y_{t-1}$ where $a = 0.026$, $b = 0.0089$. John Jackson and James Dunlevy (1982) compare Darby's procedure for defining permanent income with two alternative methods. Although they found that the three methods are empirically interchangeable, because it has the advantages of "intuitive economic appeal, computational ease, and a major gain in degrees of freedom" (p. 466), Jackson and Dunlevy recommend Darby's procedure.

TABLE 1—PARAMETER ESTIMATES OF THE GENERALIZED SPILLOVER MODEL

Process	Parameter	Estimate	Standard Error	T-Value	Probability > $\|T\|$
Static	m_r	−17.55	5.72	−3.07	0.0029
	m_y	0.19	0.06	2.98	0.0038
	y_π	−67.21	25.34	−2.65	0.0096
	y_r	−58.08	34.53	−1.68	0.0966
	y_y	1.33	0.27	4.93	0.0001
Dynamic	j	0.58	0.17	3.47	0.0009
	k	0.33	0.36	0.91	0.3651
	λ	2.21	0.98	2.25	0.0275
	μ	36.11	26.74	1.35	0.1806
	w	45.11	19.02	2.37	0.0201
Spillover	α	0.34	0.55	0.62	0.5358
	β	0.55	0.23	2.35	0.0215
Constants	s	418.85	474.60	0.88	0.3801
	h	286.37	117.55	2.44	0.0171

Survey of Current Business, and Y_N from Robert Gordon (1983). The University of Michigan Survey Research Center data for the expected Consumer Price Index, as reported by Thomas Juster and Robert Comment (1980), was used as a measure for π^e. This data has a twofold advantage. First, it provides a direct measure of inflationary expectations and thereby avoids the problems raised by Lucas (1976) associated with indirect measures; namely, the presumption that the policy regime has been stable over the sample period. Second, Nicholas Noble and Windsor Fields (1982) have demonstrated that the Michigan survey data is rational.

The parameter estimates, their standard errors, T-statistics, and the resultant level of confidence of each parameter classified as either Static (for the demand equations), Dynamic (for the rates of adjustment of demands and markets), and Spillover (degree of excess bond and commodity notional demand) are reported in Table 1.

The results reported in Table 1 are consistent with the expected signs for all of the parameters estimated. In addition, all of the Static parameters were estimated with a level of confidence of 99 percent except for y_r, which was estimated with greater than 90 percent confidence. Estimates for the Static parameters were also consistent with values previously reported by Donald Tucker (1966) and Edgar Feige (1967), who used models which specified gradual adjustment of demand and expectations, but used only partial analysis.[12]

[12] Tucker converted parameter estimates of Ronald Teigen (1964) and John Kareken and Robert Solow (1963) to a form consistent with his dynamic model. Kareken and Solow used current income and interest rates (4–6 mo. commercial paper) as independent variables in their estimates. Tucker's parameters for his desired demand functions are compared with ours below (our estimates are noted in parentheses following his):

$$g(m_r) = -3.39\ (-17.55\ S.E.\ 5.7);$$

$$f(m_y) = 0.178\ (0.19\ S.E.\ 0.06);$$

$$d(y_r) = -10.1\ (-58.08\ S.E.\ 34.5).$$

Feige specified the desired demand for money to be a function of expected income and interest rates and presented the parameters of the long-run demand equation he estimated in the form of elasticities. For comparison our estimates have been converted to elasticities. Feige's estimates were made with two measures of the money supply, M_1 and M_2:

M_1	$\beta_2(m_r) = -0.19\ (-0.044)$	$S.E.$ 0.036 (0.014)
	$\beta_1(m_y) = 1.28\ (0.933)$	$S.E.$ 0.081 (0.061)
M_2	$\beta_2 = -0.13$	$S.E.$ 0.034
	$\beta_1 = 1.07$	$S.E.$ 0.084

Feige obviously used the old M_1 data series for this work. When he estimated the equation with the more inclusive M_2, his estimates were lower. This may explain some of the difference in our estimates. Tucker specified a separate aggregate demand function for desired consumption and investment, but only investment was assumed to be a function of interest rates. Thus, his

As for the Dynamic parameters, the demand for output (j), output market (λ), and price adjustment (w) parameters were estimated with greater than 95 percent confidence, but the money demand adjustment parameter (k) and the lag in the adjustment of interest rates (μ) were not estimated with confidence. However, the estimate for k was consistent with the range of values reported by Tucker and Feige.[13]

Most importantly, the results confirmed with 98 percent confidence that spillover of excess bond demand into the commodity market, which then influences prices and output, is indeed significantly greater than zero. The hypothesis that commodity market spillover into the bond market was positive could not be confirmed with confidence.

Finally, the mean squared errors (MSE) calculated separately for the individual reduced-form equations describing Y, π, and Dr were 1330.2, 143.7, and 4.15, respectively. Although it is not clear what R^2 values mean in the context of nonlinear regressions, a 0.98 value was calculated for the Y equation, but no reasonable values for the π or Dr equations were calculated.

III. Discussion of Results

A. *The Dynamic Adjustment Parameters*

The dynamic parameters j, k, and λ are measured in units of time (in quarters), and therefore, higher (lower) values of these parameters represent slower (faster) adjustment of demand and markets. This is because the reciprocals of these parameters represent the rates at which the existing gaps between desired and actual demand and demand and supply are closed. Although these gaps would be expected to widen and narrow over time as the state of the economy changes, a "half-life" can be calculated which represents the time required for one-half of any gap to be closed by the adjustment process. The half-lives, in quarters, as well as the time for 95 percent of the gaps to be closed are reported in Table 2.

TABLE 2—DYNAMIC ADJUSTMENT PARAMETERS EXPRESSED AS HALF-LIVES

Estimate	Half-Life[a]	95 Percent Adjustment[a]
j 0.58	0.40	1.73
k 0.33	0.23	0.99
λ 2.21	1.53	6.62

[a] Quarters.

As the results in Table 2 show, money demand adjusts more rapidly than commodity demand. One-half of the adjustment of actual demand for real money balances towards the desired demand is completed in about three weeks, and 95 percent in three months. On the other hand, commodity demand adjusts toward its desired levels at a rate almost 75 percent slower than money demand. The commodity market generates output changes which remove one-half of the excess aggregate demand in about eighteen weeks, but requires over twenty months to complete 95 percent of the adjustment. This relatively rapid adjustment is probably due in large part to the significant role played by inventories in the economy.

The parameters w and μ that represent the rates at which the price level and interest rate adjust are not so easily interpreted. Because they represent the adjustment of excess demands measured in billions of dollars which, in turn, lead to changes in percentages (inflation and interest rate), their units are measured in percent/\$ billion $\times$ time. While it is clear that a larger (smaller) value implies slower (faster) adjustment, these parameter

parameter d is not equivalent to our Y_r, which includes any interest rate impact upon consumption. Tucker's $b_1 + b_2$ represent the impact of income, our y_y. However, the values for $b_1 + b_2$ reported by Tucker represented only his view of what should reasonably be expected.

$$b_1 + b_2\,(y_y) = 0.89 - 1.05\ (1.33\ (S.E.\ 0.27))$$

[13] Tucker's dynamic parameters expressing gradual adjustment of the demand for money and goods are as follows (with our estimates in parentheses):

$$j\,(j) = 0.55\ (0.58)\ S.E.\ (0.17)$$

$$m\,(k) = 0.68\ (0.33)\ S.E.\ (0.36)$$

Feige's estimates:

$$M_1 \quad \gamma(k) = 1.22, \quad S.E.\ 0.375;$$

$$M_2 \quad \gamma(k) = 1.09, \quad S.E.\ 0.399.$$

values cannot be expressed simply in units of time as can j, k, and λ.

B. *The Spillover Coefficients*

More central to the objectives of this paper is the fact that the size of the values of the spillover coefficients that we estimated imply a strong impact for monetary policy and a more complex timing relationship between the money stock and national product than is suggested by the currently popular non-market-clearing models which emphasize contract rigidity.

These results are also consistent with our earlier work. In our 1980 and 1983 papers, we investigated the implications of market spillover for the impact and timing of monetary policy using time series simulation. Our results indicated that while spillover from the commodity to bond market was not quantitatively significant, spillover from the bond market implied a strong impact on the level of economic activity. More specifically, we found (1983) that as spillover from the bond to commodity market, β, was allowed to increase from 0.0 to 0.99 monetary policy became progressively more powerful. In an experiment using numerical analysis in which $\alpha = \beta = 0.0$, and M was increased from \$157 billion to \$170 billion between the sixth and seventh quarter of the simulation experiment, Y increased from about \$550 billion to \$601 billion after nine more quarters, a 9 percent increase. When $\beta = 0.6$, the same experiment produced a 12 percent increase in Y (\$563 to 631 billion). Thus Y was \$30 billion and 5 percent higher when β was increased from 0.0 to 0.6. The peak rate of growth of Y was reached in the sixth quarter after the policy change. When β was increased from 0.0 to 0.6, the annual rate of growth of output at its peak increased from 5.2 percent to 6.2 percent.

The empirical significance of bond to commodity market spillover established here reinforces our earlier conclusions that spillover is an important element in the transmission of monetary policy and, as such, represents an additional explanation for short-run quantity adjustments.

Finally, the notion of spillover obviously suggests a more complicated linkage between the money stock and economic activity. That, *ceteris paribus*, changes in the money supply tend to produce significant short-run changes in Y in the presence of gradual adjustment of expectations at one level and gradual adjustment of output and money demand at a second level is well known as the Tucker-Laidler effect (see Donald Tucker, 1966, and David Laidler, 1968; 1982, pp. 48–57). This effect occurs because slow clearing in the money market produces exaggerated movements in interest rates, and hence, investment. When a third level of adjustment (gradual market clearing) is added, and the assumption of continually market-clearing interest rates and output is relaxed, this serves to provide the system with additional built-in stability and reduce the short-run potency of monetary policy. Interest rates need not overreact, and thus, investment expansion will not be as immediate or large.

When spillover is explicitly considered, the results above suggest that the potential for significant short-run effects upon income through monetary policy are restored. Furthermore, the source of this renewed potency is not through the money stock-interest rate channel of the Tucker-Laidler effect, but rather it is caused by the direct impact of income upon consumption and investment. The spillover of excess bond demand bypasses the interest rate mechanism and increases output directly through a quantity theory mechanism.

IV. Conclusions

In this paper we have presented reliable estimates of the parameters of a generalized disequilibrium model that expresses the short-run behavior of the U.S. economy in the face of spillover of excess demand generated by nonclearing markets. The speeds of adjustments of the financial and product markets were estimated as well as the lags in the adjustment of money and output demand toward their desired levels. The desired demand for real money balances and aggregate output were also estimated with confidence.

This work has shown that spillover must be an important consideration in the design of monetary policy. The spillover coefficient representing the size of spillover of excess bond demand into the product market was found to be large enough to significantly increase the impact of monetary policy over what might be expected in an economy without spillover.

The coefficient representing potential spillover of excess aggregate demand for commodities into the bond market could not be estimated with confidence, and therefore, the hypothesis that this type of spillover has impact upon the economy, and thus should be a factor in the design of policy, must be rejected. However, further work would seem to be indicated. If specification of the model were to allow for disaggregation of product demand into consumption and investment components, it would be possible to examine the impact of spillover in the product market more completely. This would perhaps add to our understanding of the role of inventory adjustment in timing of the business cycle and may provide a more confident and complete framework within which to prescribe monetary policy.

REFERENCES

Barro, Robert J., "Rational Expectations and the Role of Monetary Policy," *Journal of Monetary Economics*, January 1976, *2*, 1–32.

______ **and Grossman, Herschel I.**, "A General Disequilibrium Model of Income and Employment," *American Economic Review*, March 1971, *61*, 82–93.

Bergstrom, A. R., "Nonrecursive Models as Discrete Approximations to Systems of Stochastic Differential Equations," *Econometrica*, January 1966, *34*, 173–82.

Clower, Robert, "The Keynesian Counterrevolution: A Theoretical Appraisal," in F. H. Hahn and F. P. R. Breckling, eds., *The Theory of Interest Rates*, London: St. Martin's, 1965.

Darby, Michael R., "The Permanent Income Theory of Consumption—A Restatement," *Quarterly Journal of Economics*, May 1974, *88*, 228–50.

Feige, Edgar L., "Expectations and Adjustments in the Monetary Sector," *American Economic Review Proceedings*, May 1967, *57*, 462–73.

Ferguson, J. David and Hart, William R., "Liquidity Preference or Loanable Funds: Interest Rate Determination in Market Disequilibrium," *Oxford Economic Papers*, March 1980, *32*, 57–70.

______ **and** ______, "Income Determination in Market Disequilibrium: The Implications of Spillover for the Design of Monetary Policy," *Journal of Macroeconomics*, Summer 1983, *5*, 317–33.

Fischer, Stanley, "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economy*, February 1977, *85*, 191–205.

Fisher, Irving, *Appreciation and Interest*, New York: MacMillan, 1896.

______, *Theory of Interest*, New York: MacMillan, 1930.

Friedman, Milton, "A Theoretical Framework for Monetary Analysis," *Journal of Political Economy*, March/April 1970, *78*, 193–238.

Gordon, Robert J., *Macroeconomics*, 3rd ed., Boston: Little Brown, 1983.

Grossman, Herschel I., "Theories of Markets Without Recontracting," *Journal of Economic Theory*, December 1969, *1*, 476–79.

______, "Money, Interest and Prices in Market Disequilibrium," *Journal of Political Economy*, September/October 1971, *79*, 943–61.

______, "The Nature of Quantities in Market Disequilibrium," *American Economic Review*, June 1974, *64*, 509–14.

______, "The Natural-Rate Hypothesis, the Rational Expectations Hypothesis, and the Remarkable Survival of Non-Market Clearing Assumptions," *Carnegie-Rochester Conference Series on Public Policy*, Autumn 1983, Vol. 19, 225–45.

Hall, Robert E., "Is Unemployment a Macroeconomic Problem?," *American Economic Review Proceedings*, May 1983, *73*, 219–22.

Jackson, John D. and Dunlevy, James A., "The Interchangeability of Alternative Measures of Permanent Income in Regression Anal-

ysis," *Applied Economics*, October 1982, *14*, 445–68.

Juster, F. Thomas, and Comment, Robert, "A Note on the Measurement of Price Expectations," mimeo., University of Michigan Survey Research Center, 1980.

Kareken, John and Solow, Robert M., "Lags in Monetary Policy," in Commission on Money and Credit, *Stabilization Policies*, Englewood Cliifs: Prentice-Hall, 1963.

Laidler, David E. W., "The Rate of Interest and the Demand for Money—Some Empirical Evidence," *Journal of Political Economy*, December 1966, *74*, 545–55.

______, "The Permanent-Income Concept in a Macroeconomic Model," *Oxford Economic Papers*, March 1968, *20*, 11–23.

______, "The Influence of Money on Economic Activity: A Survey of Some Current Problems," in G. Clayton, J. C. Gilbert, and R. Sedwick, eds., *Monetary Theory and Policy in the 1970's*, London: Oxford University Press, 1971.

______, "Expectations, Adjustment and the Dynamic Response of Income to Policy Changes," *Journal of Money, Credit and Banking*, February 1973, *4*, 157–72.

______, *Monetarist Perspectives*. Cambridge: Harvard University Press, 1982.

Lucas, Robert E., Jr., "Econometric Policy Evaluation: A Critique," in K. Brunner and A. Meltzer, eds. *The Philips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conference Series on Public Policy, *Journal of Monetary Economics*, Suppl., 1976, 19–46.

______, "Expectations and the Neutrality of Money," *Journal of Economic Theory*, April 1972, *4*, 103–24.

______ **and Rapping, Leonard A.,** "Real Wages, Employment, and Inflation," in E. S. Phelps, ed., *Microeconomic Foundations of Employment and Inflation Theory*, New York: W. W. Norton, 1970, 257–305.

Mundlak, Yair, "Aggregation Over Time in Distributed Lag Models," *International Economic Review*, May 1961, *2*, 154–63.

Negishi, Takashi and Hahn, Frank H., "A Theorem of Nontatonnement Stability," *Econometrica*, July 1962, *30*, 463–69.

Noble, Nicholar R. and Fields, T. Windsor, "Testing the Rationality of Inflation Expectations Derived from Survey Data: A Structure-Based Approach," *Southern Economic Journal*, October 1982, *49*, 361–73.

Patinkin, Don, "The Limitations of Samuelson's Correspondence Principles," *Metroeconomica*, August 1952, *4*, 37–43.

______, *Money, Interest and Prices*, New York: Harper and Row, 1965.

Phelps, Edmund S. and Taylor, John B., "Stabilizing Powers of Monetary Policy under Rational Expectations," *Journal of Political Economy*, February 1977, *85*, 163–90.

Pippenger, John, "Monetary Policy, Homeostasis, and the Transmission Mechanism," *American Economic Review*, June 1982, *72*, 545–54.

Santomera, Anthony M. and Seater, John J., "Partial Adjustment in the Demand for Money: Theory and Empirics," *American Economic Review*, September 1981, *71*, 566–78.

Sargent, Thomas J., "A Classical Macroeconomic Model for the United States," *Journal of Political Economy*, April 1976, *84*, 207–37.

______ **and Wallace, Neil,** "Rational Expectations and the Theory of Economic Policy," *Journal of Monetary Economics*, April 1976, *2*, 169–84.

Takatoshi, Ito, "Methods of Estimation for Multimarket Disequilibrium Models," *Econometrica*, January 1980, *48*, 97–125.

Taylor, John B., "Staggered Wage Setting in a Macroeconomic Model," *American Economic Review Proceedings*, May 1979, *69*, 108–13.

Teigen, Ronald L., "Demand and Supply Functions for Money in the United States: Some Structural Estimates," *Econometrica*, October 1964, *64*, 476–509.

Tucker, Donald P., "Income Adjustments to Money-Supply Changes," *American Economic Review*, June 1966, *56*, 433–49.

______, "Macroeconomic Models and the Demand for Money under Market Disequilibrium," *Journal of Money, Credit and Banking*, February 1971, *3*, 57–83.

______, "Patinkin's Macromodel of Market Disequilibrium," *Southern Economic Journal*, October 1972, *39*, 187–203.

Federal Reserve Bulletin, various issues.

Statistical Analysis System Institute, *SAS/ETS Users Guide*, 1980.

Survey of Current Business, various issues.

The Effect of Increased Longevity on Capital Accumulation

By JONATHAN SKINNER*

One of the more startling demographic trends of the 1970's has been the sharp rise in lifespan in the United States. Between 1970 and 1980, the life expectancy of 25-year-old men and women increased by at least two years. While longevity has been steadily rising for many decades, the recent spurt has hastened the trend; the increase for both men and women in the 1970's more than doubled the improvement of the 1960's.[1] The 1970's were also characterized by a decline in labor force participation of older workers. Although rising income and increased disability payments may have explained some of the decline (Donald Parsons, 1980), it seems clear that expanding longevity has not been strongly associated with deferred retirement (Daniel Hamermesh, 1984).[2] Thus the retirement years have assumed greater importance in financial planning for the consumer as he or she retires earlier and lives longer.

The traditional life cycle model would predict greater savings rates as consumers increase accumulation of assets for their retirement. Consumption while young would drop, and there might be an increase in labor supply as individuals substitute leisure while young for greater levels of life cycle consumption. Even under a Social Security system, as long as the current generation's tax liabilities reflect their longer lifespan, total savings (private plus public) should also rise.[3]

The predicted increase in savings is substantial. A simple example using the Modigliani-Ando-Brumberg life cycle model can illustrate the extent of the savings shift. Assume that earnings and consumption are constant over working years and lifespan, respectively, and that the interest rate is zero. Consumption is then a constant proportion k of earnings, where k is the ratio of working years to one's lifespan. If consumers made plans based on average lifespan, the representative 25-year-old would have saved 18 percent of earnings in 1970 ($1-40/49$ when he plans to retire at age 65 and expects to live until age 74). In 1980, after a rise in life expectancy of two years, the representative 25-year-old would have increased savings to 21 percent of earnings, or a net rise in the savings rate (per worker) of 17 percent. That is, this simple life cycle model implies that savings in the last decade should have risen (other things held constant) by approximately 17 percent, or about 80 billion dollars per year.[4]

The evidence from aggregate time-series data appears to contradict the life cycle model's prediction of rising savings rates. The net private savings rate in the United States declined from 7.8 percent of *NNP* in 1970 to 6.2 percent in 1980. While Alan Auerbach (1982) has suggested that some of the reported decline in savings may be caused by mismeasurement, it seems clear that there has been no substantial rise in the savings rate. There are numerous explanations for

*Department of Economics, University of Virginia, Charlottesville, VA 22901. I am grateful to Maxim Engers, Daniel Hamermesh, William R. Johnson, Laurence J. Kotlikoff, John Strauss, and especially James Davies and John R. Wolfe for helpful suggestions and comments.

[1] Life table data can be found in Daniel Hamermesh (1985), the *Life Insurance Fact Book* (various years), or in unpublished form from the National Center for Health Statistics.

[2] John Wolfe (1983) provides an ingenious argument for why the Social Security benefit schedule has induced those with shorter life expectancies to retire early. We might therefore expect some rise in the average retirement age as those expecting to live longer defer retirement until age 65.

[3] If the Social Security taxes do not rise as expected lifespan increases (perhaps because current tax proceeds are paid out as current benefits, rather than retained in a trust fund), young workers will save more to offset the lower expected benefits in the future.

[4] In a two-period model, the percentage change in savings by the young is equal to the percentage change in aggregate savings.

the decline in savings, including a productivity slowdown and increasing tax burdens on capital income. The implication of this simple life cycle model is that the factors which discourage savings are substantial enough to have both reduced the savings rate *and* completely offset the savings stimulus provided by increased longevity in the 1970's.

This paper offers an explanation for why increased longevity might actually have led to a decline in the savings rate. The result hinges on a more realistic specification of consumption, bequests, and life insurance. Typically, life cycle models have included bequests motives by tacking them on to the end of the consumption function; consumers know how long they will live, to time T, and at time $T+1$, they pass along their bequests. However, there is no convincing age at which to fix time T; bequests are contingent on when the individual dies. Thus consumers have incentives to smooth contingent bequests over the life cycle, either by adjusting their consumption plans, or by buying life insurance, to insure that no matter how long (or short) they live, their descendants are provided with financial security (Christopher Pissarides, 1980). It turns out that recognizing the importance of bequests and insurance in lifetime financial plans can negate, and even reverse, the result that an increase in life expectancy increases savings. That is, the increasing longevity of the 1970's could have contributed, though marginally, to the savings slowdown of the late 1970's.

The effect of demographic changes on aggregate savings has received little attention. Dale Heien (1972) found that consumption was moderately responsive to the median age of the population, but he did not calculate the implied shift in life cycle savings. Laurence Kotlikoff (1979) suggested that increasing longevity would increase capital accumulation if workers deferred their retirement to later years. Both W. B. Arthur (1981) and Hamermesh (1985) focused on welfare implications of how increased longevity affects planned consumption, rather than on issues of capital accumulation. In addition, Hamermesh presented evidence from his survey suggesting that consumers are not only aware of current advances in life expectancies, but, to a limited extent, also project the trend forward. One reason for the paucity of research on demographic trends and savings rates may be the perception that the trends are gradual enough to exert only a negligible influence on economic variables. The recent evidence, however, suggests otherwise.

A simple model of consumption and contingent bequests is developed in the next section, and some illustrative parameters are provided to measure the effect of lifespan changes on national savings. Section II examines the role of actuarially fair life insurance policies in national savings while the conclusion is presented in Section III.

I. A Model of Contingent Bequests

This section uses a two-period model to illustrate the role of bequests in national savings. To keep the analysis simple, I assume constant earnings, interest rates, and inheritances.[5] Furthermore, consumers are unable to purchase life insurance, although this constraint will be relaxed in the next section. The general form of the utility function maximizes utility over the two possible states; surviving to the second period, or not surviving. Expected utility takes the form

$$EU=\pi U(C_1,C_2,B_2)+(1-\pi)V(C_1,B_1)$$

where U and V are utility functions, C_i is consumption at age i, π the probability of surviving to period 2, and B_i bequests at the end of age i. Specifying a utility function with separable arguments and constant elasticity of substitution among consumption and bequests, we have

$$(1)\quad EU=\begin{cases}(1-1/\sigma)^{-1}\big[C_1^{1-1/\sigma}+\delta\pi C_2^{1-1/\sigma}+\alpha(1-\pi)B_1^{1-1/\sigma}+\alpha\delta\pi B_2^{1-1/\sigma}\big] & \sigma\neq 1\\ \ln C_1+\delta\pi\ln C_2+\alpha(1-\pi)\ln B_1+\alpha\delta\pi\ln B_2 & \sigma=1\end{cases}$$

[5] See Pissarides for a general formulation of the consumption-bequest utility maximization problem. Note that when inheritances are not fixed, there will be general equilibrium effects of changes in bequests on the inheritances received by the next generation.

where σ is the intertemporal elasticity of substitution for both consumption and bequests (assumed constant), and α and δ are the bequest and time preference parameters, respectively.[6]

An alternative approach to modeling an increase in longevity would have been to add a third period, rather than to increase the probability of reaching the second. However, the observed increases in longevity are not caused by expanding frontiers of human lifespans, but by greater proportions of individuals surviving to ages 65 or 75 (James Fries, 1980). For this reason, I will adapt the two-period model in which period 1 corresponds, say, to ages 20–54, while period 2 includes ages 55–90.[7]

The Lagrangian is written

$$(2)\quad \mathscr{L} = EU + \lambda_1(Y_1 - C_1 - B_1) + \lambda_2((Y_1 - C_1)(1+r) + Y_2 - C_2 - B_2),$$

where r is the interest rate and Y_i combines earnings and accumulated inheritances received during age i. The timing of this model works as follows. First, inheritances, along with any assets from the previous period, are received at the outset of the current period. These accumulate at a rate r and, towards the end of the period are distributed along with current earnings to the recipients. Consumption is then chosen, and the remainder is passed along, either to the next period or to the next (or current) generation.

Solving for the first-order conditions, we can see immediately that $C_2 = \alpha' B_2$, where $\alpha' = \alpha^{-\sigma}$. Using the budget constraint in the second period, we can further describe second-period bequests:

$$(3)\quad B_2 = ((Y_1 - C_1)(1+r) + Y_2)/(1+\alpha').$$

Noting also that $B_1 = Y_1 - C_1$, the first-order condition $\partial\mathscr{L}/\partial C_1 = 0$ is rewritten as a function of π and C_1,

$$(4)\quad C_1^{-1/\sigma} - \alpha(1-\pi)(Y_1 - C_1)^{-1/\sigma} - \alpha\delta\pi(1+r)\left[\frac{(Y_1 - C_1)(1+r) + Y_2}{1+\alpha'}\right]^{-1/\sigma} = 0.$$

I will use (4) to examine how a change in mortality probabilities affects first-period consumption. It is then shown that, for reasonable parameters, if C_1 rises in response to longer lifespans, the aggregate capital stock will decline, and conversely.

Taking the derivative of (4) with respect to C_1 and π yields, after some substitution,

$$(5)\quad dC_1/d\pi = \frac{\alpha\sigma\left(B_1^{1-1/\sigma} - \delta(1+r)B_2^{-1/\sigma}\right)}{C_1^{\rho} + (1-\pi)\alpha B_1^{\rho} + \delta\alpha(1+r)^2\pi(1+\alpha')^{-1}B_2^{\rho}}$$

and $\rho = -1/\sigma - 1$. Note that the denominator is positive. The numerator, however, can be either positive or negative, depending on the values of B_1 and B_2. The intuition why C_1 might actually *rise* in response to an increase in longevity (thereby reducing savings while young) is to recognize that savings by the young serves a dual purpose. One is to save for retirement consumption, or the life cycle savings motive. The second is to provide bequests, either for the current or the next generation. When expected lifespan increases, the probability of leaving bequests at an early age declines, which in turn reduces the demand for contingent bequests at that

[6]An alternative to specifying bequests as the choice variable is to enter the beneficiary's utility function directly into the current utility function, or

$$EU = \pi U_1(C_1, C_2, \lambda U^*(Y^* + B_2)) + (1-\pi)U_2(C_1, \lambda U^*(Y^* + B_1))$$

where U^* is the indirect utility function for the next generation, Y^* is the descendant's income and λ is a taste parameter. Writing the descendant's indirect utility function in the same isoelastic form as equation (1) would be equivalent to defining "bequests" as $B_i + Y^*$, $i = 1,2$.

[7]It turns out that adopting this 2-period model is not a neutral choice. When one adds a third period, the implication is that consumers had previously never expected to live to the third period, thus they will always reduce consumption in the first two periods to accommodate the third period. The 2-period model is different; the idea is that the *probability* of living to the second period rises. When consumers have very little substitutability between consumption in different years (i.e., $\sigma = 0$), consumption in either period will not change at all.

age. That is, because the probability of handing over accumulated assets to one's beneficiaries has declined, there is less incentive to build up assets while young. The life cycle motive, on the other hand, works in the opposite direction; as individuals expect to live longer, they have a greater incentive to build up assets for future consumption. Which effect dominates is an empirical question, and depends on the relative bequest size of young and old individuals. If the interest rate is at least equal to the time preference rate, $B_2 > B_1$ implies that $dC_1/d\pi > 0$. That C_1 may rise in response to increased longevity therefore seems plausible, since assets of the elderly often exceed contingent bequests of the young.

I have shown that C_1 may either rise or fall as life expectancy increases. How will the change in C_1 affect aggregate capital accumulation? Let A be the steady-state ratio of capital to labor earnings (or equivalently, when earnings measure "effective labor units," the capital-labor ratio), measured at the end of the period. Then

$$A = \frac{B_1 + \pi(1+n)^{-1} B_2}{E_1 + \pi(1+n)^{-1} E_2} \tag{6}$$

where n is the rate of population growth, and E_i earnings in period i. Expressing B_2 as $(Y_1 - C_1)(1+r) + Y_2 - C_2$, and noting that from the budget constraint, $dC_2 = -[1+r)/(1+1/\alpha')]\,dC_1$, the total change in A can be expressed as

$$dA/d\pi = 1/(1+n)\bar{E} \times \left[B_2 - AE_2 - \left(1 + \frac{\pi(1+r)B_2}{B_2 + C_2}\right)\frac{dC_1}{d\pi}\right] \tag{7}$$

and $\bar{E}$ denotes aggregate earnings. The first two terms on the right-hand side are the partial effects on A of the change in the age distribution of workers, while the third term represents the change in A caused by a rise in C_1. The next task is to gather evidence about the value of assets and bequests for the young and for the elderly so that $dA/d\pi$ may be evaluated.

The numerical calculations will be made at an individual level so that we may use evidence from consumer surveys to construct the variables. Because each "period" corresponds to about 35 years of actual life, I will use average values for the first 35 years (ages 20–54) and the second 35 years (55–90). A further adjustment that must be made is to convert data drawn from yearly surveys of consumption and earnings to representative two-period values. Just as yearly consumption is the sum of daily consumption over 365 days, so also is C_1 the cumulative consumption over 35 years. I similarly define bequests in my two-period model as the stock of assets held at the end of the period. While average yearly consumption and earnings will be measured from the empirical survey, they will be adjusted appropriately in all calculations that follow.

The parameter values are presented in Table 1. Estimates of earnings, income, assets, and consumption are calculated from a sample of 16,402 families in the *Consumer Expenditure Survey* of 1972–73.[8] Two shortcomings of the asset estimates calculated from the *Survey* is that they are likely to be biased downward, and are furthermore inconsistent with the lifetime budget constraint. I therefore adjust these bequest estimates by calculating B_1 as $E - C_1$, accumulated over the 35-year first period. At the end of the first period, assets are estimated to be \$38,255 or, if asset accumulation is linear, \$19,128 on average, somewhat higher than the survey estimate of \$10,117. Similarly, the adjusted average bequest in the second period is found to be \$56,041, higher than the \$23,783 reported in the *Survey*. These adjusted values indicate, using equation (6), a capital-earnings ratio of 4.2, almost identical to the aggregate 1974 value of 4.1 (Kotlikoff

[8]Approximately 3000 families were discarded from the original sample because either they had income in excess of \$35,000 or less than \$2,000, or they failed to report earnings, income, or house values. Assets were calculated as financial wealth plus one-half the market value of the house for those under 55, and financial wealth plus the market value of the house for those over 55. Averages were weighted by the population weights provided in the *Survey*.

TABLE 1—REPRESENTATIVE LIFE CYCLE VARIABLES

Variable	Description	Value
C_1	Consumption by young	\$ 8958
E_1	After-tax earnings of young	\$10051
B_1	Bequests (assets) of young:	\$10117
	from survey adjusted	\$19128
C_2	Consumption by elderly	\$ 6127
E_2	After-tax earnings of elderly	\$ 5012
B_2	Bequests (assets) of old:	\$23783
	from survey adjusted	\$56041
δ	Time preference rate	
	(*i*) $((1.00)^{-35})$	1.000
	(*ii*) $((1.03)^{-35})$	.355
$1+r$	Interest rate $((1.04)^{35})$	3.95
π	Probability of surviving past age 54	.893
$1+n$	Population growth rate $((1.01)^{35})$	1.417
σ	Intertemporal elasticity of substitution	0.35

and Lawrence Summers, 1981; *Economic Report of the President*, 1982). There is little evidence on the value of δ, so I specify two different values, 1.0 and 0.355, corresponding to time preference rates of 0.0 and 0.03 accumulated over 35 years. A real after-tax interest rate of 0.04 seems consistent with postwar interest rate series (for example, Michael Boskin, 1978), while population growth was specified to be 1 percent this year. The variable π was estimated from 1975 life tables of the *Vital Statistics of the United States* (Table 5.7) by calculating the probability of living to age 54, conditional on reaching age 20.

There is increasing evidence that the intertemporal elasticity of substitution in consumption is considerably less than 0.5 (Robert Hall, 1981; Gilbert Ghez and Gary Becker, 1975; my dissertation, 1983; for higher estimates, see also Warren Weber, 1970, 1975); I have therefore chosen a value for σ of 0.35. Because the elasticity of substitution in bequests is unknown, I set it equal, by default, to the consumption elasticity parameter. Finally, α was calculated by evaluating $(B_2/C_2)^{\sigma}$, or 0.153.

I first consider as a benchmark case the effect of rising longevity on consumption in the pure life cycle model with no bequest motive. Setting $\alpha = 0$ and solving, I find that $dC_1/d\pi$, or the change in average yearly consumption, is $-\$999.7$ for a time preference rate equal to 0 ($\delta = 1$) and $-\$754.7$ for a time preference rate equal to .03 ($\delta = .355$).[9] Substituting into equation (9), indicates that, for the time preference rate of .03 (which will be used in most calculations), $dA/d\pi = 0.09$. That is, the capital-labor ratio responds only minimally to changes in lifespan.

How might these values for $dA/d\pi$ be interpreted in light of the increases in life expectancy during the 1970's? The probability of surviving to age 55 conditional on reaching 20 rose by 2.3 percentage points between 1970 and 1980; thus I adopt $\Delta\pi = .023$.[10] Even accepting the pure life cycle model indicates a net change in the capital labor ratio of only 0.2 percent, or holding earnings constant, a rise of about \$2 billion in the capital stock.

I next turn to calculating the change in A using the full model of consumption and bequests. Substituting from Table 1 into equation (5) indicates that, for $\delta = .355$, $dC_1/d\pi$ (once again, the change in average yearly consumption) is \$2634, while for $\delta = 1$, $dC_1/d\pi$ is \$1693. Using the lower value of δ indicates that $dA/d\pi = -0.20$, or that the increased longevity of the 1970's led to a 0.5 percent decline in the capital stock, equal to about \$19 billion in 1974. The results are not particularly sensitive to changes in the parameter values (except, as noted above, for B_1 and B_2); $dA/d\pi$ is -0.09 for $\delta = 1$, -0.08 for $\sigma = 0.6$, and -0.11 for a value of B_2, 50 percent above its original value.

Life insurance benefits comprise a substantial portion of bequests for the majority of Americans. For this reason, I now turn to a more complete model of consumption and bequests with an endogenously determined choice of actuarially fair insurance policies.

[9]When $\alpha = 0$, we can write

$$C_1 = \left(Y_1 + Y_2(1+r)^{-1}\right) / \left(1 + (\delta(1+r)\pi)^{\sigma}(1+r)^{-1}\right)$$

[10]The data is from the 1970 *Vital Statistics* and unpublished figures from the National Center for Health Statistics. The change in the probability of surviving from the midpoint of period 1 (37 years) to the midpoint of period 2 (72 years) was 6.0 percentage points; if one accepted this larger figure all estimates of capital stock change should be revised upward by 6.0/2.3.

II. Life Insurance and Bequests

This section introduces life insurance as an additional method of providing for bequests. As Pissarides has shown, the presence of actuarially fair insurance policies allows the consumer to smooth contingent bequests over the life cycle. Given the availability of these fair insurance policies, the model indicates that consumption as well as assets will rise by a trivial amount in response to an increase in longevity.[11]

Actuarially fair insurance policies may be introduced to the model as follows. When insurance is available,

$$\tilde{B}_1 = Y_1 - C_1 + \pi L \tag{8}$$

where L is the face value of the term life insurance policy and πL is the excess of the policy value over the premium paid. The first-order conditions of the new Lagrangian (see the Appendix) now include the following constraint:

$$\pi\lambda_1 - (1-\pi)(1+r)\lambda_2 = 0.$$

The consumer must choose L, C_1 and C_2, (and implicitly, $\tilde{B}_1$ and B_2), to maximize expected utility. It is shown in the Appendix that

$$dC_1/d\pi = \Delta_2\Delta_4 L/D \tag{9a}$$

$$dL/d\pi = (\Delta_1\Delta_4 - \Delta_2\Delta_3)L/D \tag{9b}$$

$$\frac{dC_2}{d\pi} = \frac{1+r}{1+\alpha}\left(L - (1-\pi)\frac{dL}{d\pi} - \frac{dC_1}{d\pi}\right) \tag{9c}$$

where $\Delta_1 = C_1^\rho + \alpha\tilde{B}_1^\rho \qquad \Delta_2 = \Delta_1 - C_1^\rho$

$$\Delta_3 = C_1^\rho + \alpha(1+r)^2\delta\beta_2^\rho/(1+\alpha')$$

$$\Delta_4 = \Delta_3 - C_1^\rho$$

$$D = \Delta_1\Delta_4(1-\pi) + \Delta_2\Delta_3\pi.$$

Because $\Delta_i > 0$, $i = 1,\dots,4$, $dC_1/d\pi > 0$, although $dL/d\pi$ may be either positive or negative. The intuition is that an increase in π provides a positive income effect, since the consumer can purchase the same L for less, holding C_1 and C_2 constant. The substitution effect, that the expected value of B_1 falls as π rises, is offset by the drop in the premium price.

To determine the magnitude of these effects, additional information from the *Life Insurance Fact Book* (1981) is used. Multiplying the average 1972 policy value of \$26,900 by the proportion of families above and below age 55, who own an insurance policy, provides estimates for insurance coverage of \$21,197 and \$18,696 for the young and old, respectively. Substituting these estimates into equations (9a)–(9c), and adding the value of the insurance policy to the assets of those over 55, the revised estimate of $dC_1/d\pi$ is a minimal \$62. The shift in the policy value, $dL/d\pi$, is an equally small $-$\$92. The change in the capital-labor ratio is written

$$\frac{dA}{d\pi} = \frac{\partial A}{\partial \pi} + \frac{\partial A}{\partial C_1}\frac{dC_1}{d\pi} + \frac{\partial A}{\partial C_2}\frac{dC_2}{d\pi} + \frac{\partial A}{\partial L}\frac{dL}{d\pi}. \tag{10}$$

Substituting parameter values again into equation (10) we have that $dA/d\pi = 0.12$. The primary reason for the positive response of assets to π is that $\partial A/\partial\pi > 0$; the compositional difference in the relative number of older people who own more assets but work less tends to increase the capital stock. Furthermore, the results are quite insensitive to what parameter values are assumed. In summary, including a realistically specified bequest motive in a life cycle model of consumption and savings indicates that, whether insurance is available or not, increasing longevity will have little or no effect on the aggregate capital stock.

III. Conclusion

Rising longevity has important implications for savings and capital accumulation. In particular, the traditional life cycle model

[11] This effect on C_1 should not be confused with Pissarides' finding that introducing fair insurance markets will increase consumption while young. In the case considered here, insurance markets are already in existence.

predicts an increase in savings rates as individuals accumulate more for retirement. In light of the accelerating trend of life expectancy, the recent decline of the savings rate appears more serious than a mere temporary stagnation. This paper has presented a more general model of consumption and bequests, and has shown that the bequest motive may negate and even reverse the positive correlation between longevity and savings implied by the pure life cycle model.

There are additional demographic factors that affect savings by influencing the demand for bequests. In particular, rising Social Security survivor benefits, a greater number of single-headed households and reduced family size are all factors likely to have reduced the necessity for bequests. Thus expressing B_1 and B_2 as functions of other demographic factors that changed during the 1970's could lead to additional theoretical support for the secular slowdown in savings.

The effects of changes in longevity on aggregate savings also depend crucially on the structure and timing of the utility-maximization problem. Expanding the model to include general equilibrium effects of shifts in bequest flows, as well as more extensive evidence on life insurance holdings by age, could modify the result that increasing lifespan has at most a negligible effect on savings. It is unlikely, for whatever parameters, that the predicted changes in savings would be of equal magnitude to that suggested by the traditional life cycle model with certain (but expanding) length of life. Both the existence of a bequest motive, and the fact that consumers response to rising lifespan as an increase in the probability of attaining old age, suggest that the rising longevity of the 1970's had little or no effect on aggregate savings rates.

APPENDIX

When actuarially fair insurance policies are available, the Lagrangian is

$$\mathscr{L} = EU - \lambda_1(Y_1 - C_1 + \pi L - \tilde{B}_1) - \lambda_2((Y_1 - C_1 - (1-\pi)L)(1+r) + Y_2 - C_2 - B_2).$$

It is straightforward to show that

$$\text{(A1)} \quad C_1^{-1/\sigma} - \alpha\tilde{B}_1^{-1/\sigma} = 0$$

$$\text{(A2)} \quad C_1^{-1/\sigma} - \alpha(1+r)\,\delta B_2^{-1/\sigma} = 0.$$

Substituting in for $\tilde{B}_1$ and B_2 (and noting that $C_2 = \alpha' B_2$), and taking the total derivative of (A1) and (A2), we have

$$\text{(A3)} \quad -(C_1^{\rho} + \alpha\tilde{B}_1^{\rho})\,dC_1 + \alpha\tilde{B}_1^{\rho}(L\,d\pi + \pi\,dL) = 0$$

$$\text{(A4)} \quad -(C_1^{\rho} + \Delta_4)\,dC_1 + \Delta_4(L\,d\pi - (1-\pi)\,dL) = 0$$

and $$\Delta_4 = \alpha(1+r)^2\,\delta B_2^{\rho}/(1+\alpha').$$

From (A3) and (A4), it is straightforward to derive the expressions in the text.

In a model with insurance, A is defined somewhat differently;

$$A = \Big[Y_1 - C_1 + \pi(1+r)(1+n)^{-1} \times (Y_1 - C_1 - (1-\pi)L) + \pi(1+n)^{-1}(Y_2 - C_2)\Big] \Big/ \Big[E_1 + \pi(1+n)^{-1}E_2\Big].$$

Note that the premiums collected in the first period still comprise a portion of total assets (since they haven't been paid out), so that assets of the young are $Y_1 - C_1$.

REFERENCES

Arthur, W. B. "The Economics of Risks to Life," *American Economic Review*, March 1981, *71*, 54–64.

Auerbach, Alan J., "Issues in the Measurement and Determination of Business Savings," Working Paper No. 1024, National Bureau of Economic Research, November 1982.

Boskin, Michael J., "Taxation, Savings, and the Rate of Interest," *Journal of Political and Economy*, April 1978, *86*, S3–28.

Davies, James B., "Uncertain Lifetime, Consumption, and Dissavings in Retirement," *Journal of Political Economy*, June 1981, *89*, 561–77.

Farber, Henry S., "Individual Preferences and Union Wage Determination: The Case of United Mine Workers," *Journal of Political Economy*, October 1978, *86*, 923–42.

Fries, James F., "Aging, Natural Death, and the Compression of Morbidity," *New England Journal of Medicine*, July 17, 1980, 130–35.

Gersovitz, Mark, "Savings and Nutrition at Low Incomes," *Journal of Political Economy*, October 1983, *91*, 841–55.

Ghez, Gilbert and Becker, Gary S., *The Allocation of Time and Goods over the Life Cycle*, New York: Columbia University Press, 1975.

Hall, Robert E., "Intertemporal Substitution in Consumption," Working paper No. 720, National Bureau of Economic Research, July 1981.

Hamermesh, Daniel S., "Expectations, Life Expectancy and Economic Behavior," *Quarterly Journal of Economics*, forthcoming 1985.

______, "Life-Cycle Effects on Consumption and Retirement," *Journal of Labor Economics*, July 1984, *2*, 1–7.

Heien, Dale M., "Demographic Effects and the Multiperiod Consumption Function," *Journal of Political Economy*, January/February 1972, *80*, 125–38.

Kotlikoff, Laurence J., "Some Economic Implications of Life Span Extensions," Working Paper No. 155, UCLA, May 1979.

______ **and Summers, Lawrence H.**, "The Role of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy*, August 1981, *89*, 706–32.

Menchik, Paul and David, Martin, "Income Distribution, Lifetime Savings, and Bequests, *American Economic Review*, September 1983, *73*, 672–90.

Parsons, Donald, "The Decline in Male Labor Force Participation," *Journal of Political Economy*, February 1980, *79*, 117–34.

Pissarides, C. A., "The Wealth-Age Relation with Life Insurance," *Economica*, November 1980, *47*, 451–58.

Skinner, Jonathan, "The Effects of Variable Lifespan and Uncertain Earnings on Consumption: A Theoretical and Empirical Analysis," unpublished doctoral dissertation, UCLA, 1983.

Weber, Warren E., "The Effect of Interest Rates on Aggregate Consumption," *American Economic Review*, September 1970, *60*, 591–600.

______, "Interest Rates, Inflation, and Consumer Expenditures," *American Economic Review*, December 1975, *65*, 843–58.

Wolfe, John R., "Perceived Longevity and Early Retirement," *Review of Economics and Statistics*, November 1983, *65*, 544–51.

Yaari, Menahem E., "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer," *Review of Economic Studies*, April 1965, *32*, 137–50.

American Council of Life Insurance, *Life Insurance Fact Book*, 1981 and various years.

National Center for Health Statistics, *Vital Statistics of the United States*, Washington, 1970, 1975.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington, 1982.

Purchasing Power Parity Did Not Collapse During the 1970's

By NURHAN DAVUTYAN AND JOHN PIPPENGER*

When data from the 1970's are used to evaluate purchasing power parity, the results generally lead to rejection. See, for example, Paul Krugman (1978), Richard Roll (1980), Jacob Frenkel (1981), and Robert Cumby and Maurice Obstfeld (1982). Probably the best known study is by Frenkel where he compares the performance of purchasing power parity (*PPP*) during the 1920's and 1970's, and concludes that the theory worked well in the 1920's, but not during the 1970's. We believe that purchasing power parity worked about as well in the 1970's as it did in the 1920's. What was different between the two periods was that monetary policy was more coordinated in the 1970's than it was in the 1920's. It is the coordination of monetary policy, not the collapse of *PPP*, that causes conventional tests to reject *PPP* for the 1970's. In addition, we believe that, contrary to conventional wisdom, *PPP* works at least as well under monetary stability as during inflation.

We develop our argument as follows. Section I summarizes the *PPP* equilibrium based on commodity arbitrage and examines how *PPP* can fail. Section II uses this analysis to reinterpret the evidence from the 1920's and 1970's, and Section III discusses the interpretation of the errors from *PPP*.

I. Theoretical Background

A. *Commodity Arbitrage*

Suppose that there are no transaction costs or other impediments to trade. All goods are tradable and effective arbitrage implies the relative version of purchasing power parity.

$$(1)\qquad \frac{(\$/£)_t}{(\$/£)_0} = \frac{(\$/Q)_t/(\$/Q)_0}{(£/Q)_t/(£/Q)_0},$$

where Q is the same bundle of goods for both countries. For simplicity, we define R_t and P_t as follows:

$$(\$/£)_t/(\$/£)_0 \equiv R_t$$

and

$$\frac{(\$/Q)_t/(\$/Q)_0}{(£/Q)_t/(£/Q)_0} \equiv P_t.$$

Now equation (2) describes the relative version of *PPP*:

$$(2)\qquad R_t = P_t$$

In most empirical studies, the test equation is

$$(3)\qquad \log R_t = \alpha_0 + \alpha_1 \log \hat{P}_t,$$

where $\hat{P}_t$ is based on consumer or wholesale indexes, or *GNP* deflators. In many cases R_t does not include the base-period exchange rate and α_0 is an estimate of the log of the base exchange rate.

B. *Why PPP Fails*

From an arbitrage perspective, there are essentially two reasons for deviations from equation (2). Arbitragers may not respond to profitable opportunities, or transaction costs and other impediments may inhibit trade. The first explanation seems unlikely, so in the first three subsections we concentrate on the effects of transaction costs. The following two subsections discuss why estimates of equation (3) can fail even when (2) holds.

1. *Nontradables*: With zero transaction costs, all goods are tradable. When we drop that

*Visiting Lecturer and Professor of Economics, respectively, University of California, Santa Barbara, CA 93106. We thank the following individuals for their comments: John Mussachia, Jürg Niehans, an anonymous referee, and various members of the research staff at the Federal Reserve Bank of San Francisco.

assumption, it is convenient to divide goods into two groups: tradables with zero transaction costs and nontradables with high transaction costs. In the absence of trade restrictions, arbitrage keeps relative prices between tradables equal across countries, but this linkage breaks down for nontradables and relative prices between nontradables as well as between tradables and nontradables can differ between countries. As a result, when there are real shocks, equation (2) holds for tradables but not for nontradables.

Although it is generally recognized that real shocks introduce errors into *PPP* and reduce the R^2, it is not generally appreciated that these shocks introduce measurement error and bias the estimate of α_1 toward zero. As an example, suppose the variance in P, σ_p^2, comes from two independent sources. Monetary shocks that do not alter relative prices, σ_M^2, are one source. For this source, equation (2) holds. In addition, there are real shocks that change relative prices between nontraded goods, σ_R^2. For this component, there is no systematic relation between price levels and exchange rates.[1]

$$\sigma_P^2 = \sigma_M^2 + \sigma_R^2$$

Under these conditions, ordinary least squares yields the following estimate for α_1:[2]

$$\text{(4)} \qquad \operatorname{plim} \hat{\alpha}_1 = 1.0/\left[1.0 + \left(\sigma_R^2/\sigma_M^2\right)\right].$$

As inflationary shocks dominate real shocks, estimates of α_1 approach unity. As monetary shocks disappear, R^2 declines and estimates of α_1 approach zero even though equation (2) holds perfectly for monetary shocks and real shocks have not increased.

[1]As Bela Belassa (1964), Frenkel, and others have recognized, changes in relative prices between tradables and nontradables also introduce errors into *PPP*. However, the econometric implications for such changes are much more difficult to determine because there is a systematic relation between these changes and the exchange rate. An increase in the price of nontradables relative to tradables tends to be associated with appreciation. How this change is related to P depends on the weights in the price indexes.

[2]For an excellent discussion of the effects of measurement error, see Albert Madansky (1976).

2. *Tradables*: The assumption of zero transaction costs and no trade restrictions for tradables is analytically convenient, but not very accurate. Relative prices among tradables also can diverge between countries. As in the case of nontradables, this divergence introduces error into *PPP* and reduces both the R^2 and regression coefficient. The primary difference between tradables and nontradables is that differences in relative prices are limited for tradables, but not for nontradables.

This distinction is important because the error structure is different in the two cases. If errors in *PPP* are due primarily to changes in relative prices between tradables, the errors are bounded and real exchange rates should not behave like a martingale. If errors in *PPP* are due to changes in relative prices between nontradables, or between tradables and nontradables, no such restriction applies and real exchange rates can behave like a martingale. A given shock can alter the relative price between nontradables and the next shock may be as likely to reinforce as to offset the initial shock.

3. *The Asset Approach*: Asset models of exchange rates like Rudiger Dornbusch (1976) have become popular because they appear to explain the volatility of exchange rates. An essential feature of these models is that commodity markets adjust slowly relative to markets for assets. As a result, commodities are nontradable in the short run and tradable in the long run. If this approach is correct, the errors in equation (3) should be relatively small in the long run and real exchange rates should not behave like a martingale.

4. *Simultaneity*: Even if equation (2) held exactly, equation (3) could yield estimates of α_1 that are not equal to unity. Neither price levels nor exchange rates are exogenous variables, which raises the possibility of bias due to simultaneous equations. See Krugman for an example of how simultaneity can bias estimates of α_1 toward zero.

5. *Unequal Weights*: In equation (2), the weights in the price indexes are the same for

TABLE 1—MONTHLY *GLS* ESTIMATES OF EQUATION (3) USING THE DOLLAR PRICE OF FOREIGN EXCHANGE AND WHOLESALE INDEXES

Country and Period	α_0	α_1	R^2/ Standard Error	Durbin-Watson	ρ
France	0.636	1.357	0.536	1.57	0.814
1920–25	(0.137)	(0.151)	0.056		
Germany	−0.746	0.976	0.997	2.02	0.318
1919–24	(0.082)	(0.006)	0.366		
England	−0.043	0.711	0.489	1.39	0.761
1920–March 1925	(0.018)	(0.093)	0.025		
Canada	0.048	0.289	0.267	1.52	0.824
1919–25	(0.07)	(0.053)	0.011		
Japan	0.036	0.201	0.204	1.23	0.957
1919–25	(0.033)	(0.044)	0.015		

Note: All data are from various issues of the *Federal Reserve Bulletin* except for German wholesale prices, which are from *European Currency and Finance* (J. P. Young, 1925). The base period is 1913 and the base exchange rate is mint par. These regressions use the SAS AUTOREG procedure with one-period serial correlation for the error term. Standard errors are shown in parentheses.

both countries. Using *CPI* or wholesale price indexes or *GNP* deflators violates this requirement and introduces measurement error. Consider the following simple example. Country *A* produces only wheat and country *B* only cloth. Some real shock causes the wheat price of cloth to rise proportionately in both countries, but the exchange rate does not change and neither does the ratio of price levels P_t when the weights are the same. Now replace P_t with $\hat{P}_t$, the *GNP* deflator for *A* divided by the *GNP* deflator for *B*. $\hat{P}_t$ falls when P_t and the exchange rate are unchanged. This movement in $\hat{P}_t$ biases the estimate of α_1 toward zero and reduces the R^2 even if equation (2) holds exactly.

II. The Evidence

A. *Inflation vs. Monetary Stability*

Table 1 shows monthly *GLS* estimates of equation (3) during the 1920's for the United States vs. five other countries. The R^2 and estimate of the regression coefficient support the conventional wisdom that purchasing power parity works well under inflationary conditions, but not under more normal conditions. The R^2 is relatively high and the estimate of α_1 fairly close to unity for France and Germany, but for Canada and Japan the R^2 is low and the estimate of α_1 is over 10 standard errors below unity.[3]

Subsection 5 above suggests that the R^2 and regression coefficient provide a poor basis for judging *PPP*. The real exchange rates shown in Table 2 illustrate that point. Table 2 uses the same data as Table 1, but tells a very different story. For France, the best year yields a predictive error of over 13 percent. For Germany the best *PPP* can do is about a 25 percent error, and that is after the hyperinflation is over.

The performance of *PPP* for the noninflationary countries is much better. For the worst years, *PPP* yields a predictive error of 11, 7, and 17 percent for Canada, England, and Japan, respectively. When *PPP* predicts exchange rates in 1924 based on rates in 1913, for these three countries the theory yields a maximum predictive error of about 9 percent and an average predictive error of about 1 percent.

The information in Table 2 is not totally absent from Table 1, but one must look

[3]In terms of relative monetary stability, England lies between these two pairs of countries. From 1919 to 1926, wholesale prices in Canada, Japan, and the United States all fell about 25 percent. In England, prices fell by about 35 percent.

TABLE 2—ACTUAL DOLLAR PRICES OF FOREIGN EXCHANGE AS A PROPORTION OF THE PARITY RATE BASED ON WHOLESALE INDEXES AND 1913 MINT PAR AS A BASE-PERIOD EXCHANGE RATE

Year	France	Germany	Canada	England	Japan
1919	–	0.464	0.946	–	1.137
1920	0.782	0.449	0.938	0.994	1.010
1921	0.868	0.563	1.008	1.070	1.173
1922	0.863	0.534	0.933	0.972	1.113
1923	0.750	0.665[a]	0.896	0.970	1.114
1924	0.758	0.748	0.911	1.007	1.041
1925	0.716	–	0.912	–	0.972

Note: Annual averages of monthly data used in Table 1.
[a] October was omitted. If that month is included the figure is 1.638.

carefully. Since Table 1 uses a base-period exchange rate, the large negative intercept for Germany is a sign that *PPP* is seriously biased. An examination of the standard errors also suggests that equation (3) did not work as well for France and Germany as for England, Canada, and Japan. For the first two countries, the standard errors of the regression are 0.056 and 0.366, but for the other three countries they are 0.025, 0.011, and 0.015.

A comparison of Tables 1 and 2 provides one example of how dangerous it can be to evaluate *PPP* solely on the basis of R^2 and regression coefficients. The conclusion that purchasing power parity failed in the 1970's provides another example.

B. *Frenkel's Evidence*

Table 3 replicates Frenkel's estimates for the 1920's and Table 4 shows his results for the 1970's. The average standard error for the 1920's is 0.102. For the 1970's, it is only 0.029. If we drop the dollar/franc rate, which appears only in the 1920's, the average standard error for the 1920's rises to 0.122. Based on the predictive error, this evidence does not support the claim that *PPP* collapsed in the 1970s.

Frenkel attributes the decline in *PPP* to "...the volatile character of the 1970s which witnessed great turbulence in the world economy and large volumes of real shocks like the oil embargo, supply shocks, commodity booms and shortages, shifts in the demands for money and differential productivity growth" (p. 162). However, the evidence presented here, and a more balanced view of history, suggest that the 1970's were no more turbulent than the 1920's. A comparison of the standard errors for *PPP* in the two periods certainly suggests that real shocks did not interfere with *PPP* any more in the 1970's than in the 1920's. In addition, recent turbulence such as the oil shock is fresh in our mind, but we tend to forget that the 1920's were also very turbulent.[4] In addition to hyperinflations, there were the uncertainties associated with the reestablishment of trade after World War I, the various reparations crises, and, perhaps most important, the massive worldwide agricultural depression. From 1919 to 1921, as the U.S. wholesale price index fell by 29 percent, U.S. farm prices fell 30 percent relative to industrial prices. From 1921 to 1925, with a stable price level, farm prices rose 25 percent relative to industrial prices.[5]

C. *Another Look at the Recent Float*

Table 5 shows monthly estimates of equation (3) for the United States and Canada for 1972 through 1977, and 1978 through early

[4] Frenkel recognized the effect of changes in relative prices between tradables and nontradables. Indeed, he based his conclusions to some extent on an estimate of the change in relative prices using the ratio of consumer to wholesale indexes. Unfortunately, he only did these computations for the 1970's, so he could not compare their behavior in the two periods.

[5] See *Historical Statistics of the United States...*, Vol. I (1975, p. 199).

TABLE 3—PURCHASING POWER PARITIES: INSTRUMENTAL VARIABLES, MONTHLY DATA DURING THE 1920's[a]

Dependent Variable $\ln S_t$	Price Index	Constant	$\text{Ln } P_t$	Standard Error	Durbin-Watson	ρ
Mark/Pound (Feb. 1921 –Aug. 1923)	Wholesale	−1.676 (0.178)	1.026 (0.017)	0.221	2.01	0.24
	Cost of Living	−1.575 (0.423)	1.084 (0.041)	0.367	2.06	0.50
Franc/Pound (Feb. 1921 –May 1925)	Wholesale	0.562 (0.207)	1.141 (0.064)	0.044	1.82	0.53
	Material	0.613 (0.180)	1.081 (0.054)	0.042	2.18	0.48
Dollar/Pound (Feb. 1921 –May 1925)	Wholesale	−0.118 (0.482)	0.897 (0.267)	0.019	1.99	0.85
	Material	−0.073 (0.453)	0.847 (0.245)	0.022	1.83	0.80
Franc/Dollar (Feb. 1921 –May 1925)	Wholesale	1.183 (0.157)	1.091 (0.109)	0.054	1.70	0.58
	Material	1.243 (0.130)	0.992 (0.085)	0.050	1.74	0.54

Source: Frenkel, Table 1, p. 148 (reprinted with permission of the *European Economic Review*).
Note: S_t is the spot rate, not the spot rate relative to a base-period rate. As a result, the intercept serves as an estimate of the log of the rate in the base period.
[a]Standard errors are shown in parentheses.

TABLE 4—PURCHASING POWER PARITIES: INSTRUMENTAL VARIABLES, MONTHLY DATA: JUNE 1973–JULY 1979[a]

Dependent Variable $\ln S_t$	Constant	$\ln P_t^c$	$\ln P_t^w$	Standard Error	Durbin-Watson	ρ
Dollar/Pound	0.712 (0.149)	0.165 (0.507)		0.027	1.63	0.963
	2.982 (2.978)		1.07 (0.897)	0.029	1.66	0.998
Dollar/Franc	−1.521 (0.027)	0.184 (0.374)		0.029	2.26	0.863
	−1.57 (0.047)		−1.07 (0.817)	0.029	2.3	0.901
Dollar/DM	−0.900 (0.018)	1.786 (0.230)		0.034	1.69	0.739
	−0.908 (0.175)		2.217 (0.263)	0.031	1.96	0.759

Source: Frenkel, Table 2, p. 149 (reprinted with permission of the *European Economic Review*).
Note: P_t^c and P_t^w are, respectively, the consumer and wholesale relative price levels. Since S_t is the spot rate, not the spot rate relative to a base period rate, the intercept is an estimate of the log of the rate in the base period.
[a]Standard errors are shown in parentheses.

TABLE 5—MONTHLY *GLS* ESTIMATES OF EQUATION (3) FOR THE UNITED STATES VS. CANADA, ARGENTINA, BRAZIL, AND ISRAEL: WHOLESALE INDEXES

Country	Period	α_0	α_1	R^2/ Standard Error	Durbin-Watson	ρ
Canada	Jan. 1972	−0.02	0.25	0.03		
	Dec. 1977	(0.01)	(0.16)	0.010	1.07	0.82
	Jan. 1978	−0.15	0.82	0.37		
	Feb. 1984	(0.00)	(0.12)	0.010	1.62	0.73
Argentina	June 1973	1.11	1.12	0.89		
	July 1979	(0.17)	(0.04)	0.134	2.15	0.83
Brazil	June 1973	−3.82	0.98	0.96		
	July 1979	(0.03)	(0.02)	0.018	1.61	0.81
Israel	June 1973	−1.82	1.11	0.88		
	July 1979	(0.08)	(0.04)	0.046	2.09	0.81

Note: IFS data tape October 1984 with no base-period exchange rate. These regressions use the SAS AUTOREG procedure with one-period serial correlation for the error term. Standard errors are shown in parentheses.

1984. In terms of the R^2 and regression coefficient, *PPP* fails for the earlier period but not for the later period.[6] The reason for the difference is that from 1972 through 1977 wholesale prices in Canada rose only 5 percent more than in the United States, while from 1978 through early 1984 the difference was over 14 percent. The divergence in monetary policy also caused the variance in the log of relative price levels to increase by about four times in the second period.

Although by conventional standards *PPP* appears to fail for Canada during the early 1970's, in terms of the standard error of the equation, *PPP* performs as well during this period as from 1978 to early 1984. In addition, from that perspective, *PPP* performs better for Canada during the 1970's than it does for all other countries in the 1920's.

Table 5 also shows *GLS* estimates of equation (3) between the United States and three inflationary countries: Argentina, Brazil and Israel. Purchasing power parity worked as well for these countries in the 1970's as it did for France and Germany in the 1920's. Given the evidence for Canada, Argentina, Brazil, and Israel, it is impossible to conclude that *PPP* worked well in the 1920's, but collapsed in the 1970's.

[6]For the second period, other programs yield a different coefficient for the same *rho*. Hildreth-Lu, for example, gives 0.68 (0.12) as an estimate of α_1, which is also a substantial improvement over the earlier period. The standard error of the regression is essentially unchanged.

III. Interpretation of Errors

Using an arbitrage approach, the theoretical section identified five sources for the failure of purchasing power parity. What can we say about the relative importance of these sources?

A. *Different Weights*

Recent research by John Mussachia (1984) using indexes with identical weights suggests that, except perhaps for very stable monetary episodes, different weights in the price indexes are not a major source for the observed errors in purchasing power parity.[7]

B. *Assets and Tradables*

Asset models like Dornbusch's and changes in relative prices between tradables imply that movements in the real exchange rate are bounded. A number of studies, however, have found that the real exchange rate essentially performs a martingale. See Roll, Michael Adler and Bruce Lehman (1983), Michael

[7]Of course, if *PPP* is used as a guide to policy, even small errors due to the wrong weights can have serious consequences.

Darby (1980), and Pippenger (1982).[8] Since a martingale implies that real exchange rates are unbounded, a martingale is inconsistent with asset models like Dornbusch's and changes in relative prices between tradables.[9]

C. *Simultaneity*

Ordinary least squares estimates of *PPP* are subject to bias due to simultaneous equations, but this bias does not appear to be a major source for the low R^2 and estimates of α_1. In Tables 3 and 4, Frenkel used two-stage least squares with instrumental variables and *PPP* still appeared to fail when relative price levels moved together. In addition, the results of our earlier paper (1984) suggest that, once the effects of measurement error are reduced substantially, the instrumental variable technique used by Frenkel and others has little affect on estimates of α_1.

D. *Nontradables*

Given the evidence as it now stands, the most reasonable interpretation appears to be that the errors from *PPP* are dominated by changes in relative prices for nontradables, which is consistent with a martingale.[10] If this interpretation is correct, then the relatively large errors for inflationary countries suggests that monetary instability increases the movement of relative prices for nontradables.

IV. Summary and Conclusions

The evidence suggests that the errors in purchasing power parity are dominated by shocks that alter relative prices for nontradables. Although it is widely recognized that such shocks introduce errors into *PPP*, it is not generally appreciated that they also can bias the estimate of α_1 toward zero. As a result, *PPP* can predict well for periods of relative monetary stability even when estimates of R^2 and the regression coefficient are close to zero.

In addition, an R^2 and estimate of α_1 close to unity for inflationary episodes does not imply that *PPP* works better under inflationary conditions than it does under more stable monetary policy. These estimates simply imply that purely monetary shocks are swamping the effects of real shocks. Indeed, the predictive error from purchasing power parity tends to be larger with inflation.

When Frenkel and others compared the results of testing *PPP* for the 1920's and 1970's, they misinterpreted the sharp declines in the R^2 and regression coefficient. They attributed this decline to a collapse of *PPP*, when in fact it reflected a reduction in the relative importance of monetary shocks. If there has been a decline in the predictive ability of *PPP* from the 1920's to the 1970's, it has been very modest. Purchasing power parity did not collapse in the 1970's.

[8]These studies tend to concentrate on the 1970's. However, our analysis of the 1920's yields similar results. Identically, weighted indexes also yield the same results. See Mussachia.

[9]There is, however, a question about the power of all the standard tests to identify a stochastic process with weak elastic reflecting barriers.

[10]If this interpretation is correct, the strength of the dollar during 1983 and 1984 is primarily the result of an increased demand for nontradables in the United States that has lowered the price of tradables relative to nontradables.

REFERENCES

Adler, Michael and Lehmann, Bruce, "Deviations from Purchasing Power Parity in the Long Run," *Journal of Finance*, December 1983, *38*, 1471–87.

Belassa, Bela, "The Purchasing-Power Parity Doctrine: A Reappraisal," *Journal of Political Economy*, December 1964, *72*, 584–96.

Cumby, Robert E. and Obstfeld, Maurice, "International Interest Rate and Price-Level Linkages Under Flexible Exchange Rates: A Review of Recent Evidence," Working Paper No. 142, Department of Economics, Columbia University, April 1982.

Darby, Michael R., "Does Purchasing Power Parity Work?," Working Paper No. 607, National Bureau of Economic Research, December 1980.

Davutyan, Nurhan and Pippenger, John, "Testing Purchasing Power Parity," manuscript,

September 1984.

Dornbusch, Rudiger, "Expectations and Exchange Rate Dynamics," *Journal of Political Economy*, December 1976, *84*, 1161–76.

Frenkel, Jacob A., "The Collapse of *PPP* During the 1970's," *European Economic Review*, May 1981, *16*, 145–65.

Krugman, Paul, "Purchasing Power Parity and Exchange Rates: Another Look at the Evidence," *Journal of International Economics*, August 1978, *8*, 397–407.

Madansky, Albert, *Foundations of Econometrics*, Amsterdam: North-Holland, 1976.

Mussachia, John, "A Reexamination of the Purchasing Power Parity Theory During the Recent Floating Rate Period," unpublished doctoral dissertation, University of California-Santa Barbara, 1984.

Pippenger, John, "Purchasing Power Parity: An Analysis of Predictive Error," *Canadian Journal of Economics*, May 1982, *15*, 335–46.

Roll, Richard, "Violations of *PPP* and their Implications for Efficient International Commodity Markets," in M. Sarnat and G. P. Szego, eds., *International Trade and Finance*, Vol. I, Cambridge: Ballinger, 1980, 133–76.

Young, John P., *European Currency and Finance*, Commission of Gold and Silver Inquiry, U.S. Senate, Washington: USGPO, 1925.

U.S. Department of Commerce, Bureau of Census, *Historical Statistics of the United States: Colonial Times to 1970, Bicentennial Edition*, Washington: USGPO, 1975.

Speeding, Coordination, and the 55 MPH Limit

By CHARLES A. LAVE*

Do laws coordinate or restrain? A number of recent papers discuss the optimality of the 55 mph national maximum speed limit (NMSL), and evaluate the tradeoff of time-lost vs. lives-saved resulting from the lowered speed (James Jondrow et al., 1983; Dana Kamerud, 1983; Thomas Forester et al., 1984). These papers all implicitly accept the conventional wisdom—speed kills, slower is safer.

This conventional wisdom leads to laws designed as *limits* on behavior, whereas "...the crucial element is often coordination. People need to do the right things at the right time in relation to what others are doing" (Thomas Schelling, 1978, p. 121). There are indeed some traffic laws that establish conventions of expected conduct: we ask that motorists drive to the right, not because driving on the left is evil, but because it is important that the direction of flow be commonly agreed upon. Likewise, traffic lights are best viewed as a coordinating device: allowing free flow to alternating lanes of traffic to reduce the confusion and loss of time in unsignalized intersections.

For peculiar historical reasons, speed laws evolved as *limits* on driver behavior, rather than as signaling devices meant to *coordinate* it. Guided by the limit-rationale, police concentrate on those drivers who exceed the legal speed, and tend to ignore those drivers who disrupt coordination by traveling much slower than the norm.

This paper tests these differing views of the law by examining the current effects of the 55 mph NMSL—should it be viewed as a coordinating mechanism or a limiting mechanism? I measure the effects of limit-defying behavior (speeding), and absence of coordination (speed variance) on the fatality rate. Based on analysis of 1981 and 1982 state cross-section data, I find that there is no statistically discernable relationship between the fatality rate and average speed, though there is a strong relationship to speed variance. When most cars are traveling at about the same speed, whether it is a high speed or a low one, the fatality rate will be low—presumably because the probability of collision will be low. Variance kills, not speed.

I. Data and Methodology

The dependent variable is fatalities per 100 million vehicle miles traveled; data points are state averages. Since fatality rates differ by type of highway, I looked separately at six different types of high-speed roads: rural interstates, arterials, and collectors; and urban freeways, interstates, and arterials.[1] Separate regressions were fitted to data for 1981 and 1982; thus there were twelve distinct equations—six highway types for each of two years. I screened out any data point based on five or fewer fatalities; and Alaska and Hawaii were excluded because of their markedly atypical highways and driving conditions.

Table 1 shows the characteristics of each subset of the data. ("Speed Variance" is a measure of the dispersion of speeds among drivers. The distribution is approximately bell shaped: the "Average Speed" is about at the center of this distribution, and the "85th Percentile Speed" is about one standard de-

*Department of Economics, University of California, Irvine, CA 92717. Amihai Glazer made a major contribution to the conceptualization of the research; he is innocent of any errors in its execution. The work was done while I was serving as a member of the National Research Council's Commission for the Study of Benefits and Costs of the 55 MPH National Maximum Speed Limit; though the conclusions do not necessarily represent those of the staff and members of the Commission.

[1] "Urban freeways" also include urban expressways, hence their average standard is lower than interstates. "Arterials" have improved shoulders and wider lanes than "collectors."

TABLE 1—MEANS OF THE VARIABLES

Highway Type	*Fatality Rate*[a]	*Average Speed* (mph)	*Percent of Cars > 55 mph*	*Percent of Cars > 65 mph*	*85th % Speed* (mph)	*Speed Variance*[b] (85 – Avg)	*Citations per Driver*[c]	*N*
1981								
Rural Interstate	1.81	58.1	68.7	9.31	63.2	5.13	74.0	41
Rural Arterial	4.97	54.1	43.4	4.37	60.0	5.85	72.0	46
Rural Collector	4.11	51.7	33.4	4.42	58.7	6.99	75.3	41
Urban Freeway	3.25	54.9	46.3	2.94	59.9	5.00	58.2	19
Urban Interstate	1.37	55.8	54.1	4.46	61.1	5.33	59.4	26
Urban Arterial	2.67	51.9	31.2	2.20	58.1	6.12	63.7	23
1982								
Rural Interstate	1.50	59.0	73.1	14.2	65.1	6.12	67.0	44
Rural Arterial	4.24	54.4	47.1	6.23	61.2	6.81	64.8	47
Rural Collector	4.32	51.8	35.4	5.54	59.6	7.85	67.3	41
Urban Freeway	1.77	56.2	55.1	6.80	62.5	6.32	53.9	18
Urban Interstate	1.24	56.6	61.8	8.53	63.0	6.41	55.2	27
Urban Arterial	2.33	52.2	35.9	4.02	59.4	7.25	55.7	21

Source: Highway Statistics, U.S. Department of Transportation.
[a]Fatalities per 100 million vehicle miles traveled.
[b]85th percentile speed minus the average speed: a rough measure of the standard deviation of the distribution of speeds.
[c]Speeding citations (on all highway types) per 100 drivers, per year. Variations in the average, across highway types, occur because state-composition varies across subsamples.

TABLE 2—RANGE OF VARIATION AMONG STATES

Variables	Mean	Lowest Value	Highest Value
Rural Interstates			
Fatality Rate	1.81	0.39	4.79
Average Speed	58.1	54.8	62.5
Percent of Drivers > 55 mph	68.7	40.3	88.8
Percent of Drivers > 65 mph	9.31	1.70	28.6
85th % Speed	63.2	58.8	69.4
85th Percentile – Average Speed	5.13	2.70	9.10
Citations per Driver	74.0	24.1	193.
Urban Freeways			
Fatality Rate	3.25	0.89	15.5
Average Speed	54.9	51.1	57.2
Percent of Drivers > 55 mph	46.3	23.1	64.6
Percent of Drivers > 65 mph	2.94	0.40	8.10
85th % Speed	59.9	57.4	63.8
85th Percentile – Average Speed	5.00	1.80	7.50
Citations per Driver	58.2	24.1	118.

viation to the right. Thus, the measure "85th Percentile – Average Speed" is a proxy for the standard deviation of observed speeds.)

Driving speeds and fatality rates differ considerably among states. Table 2 illustrates this variation for two subsets of the data: interstate rural roads and urban freeways in 1981.

II. A Model of the Fatality Rate

On a priori grounds, we can say that the fatality rate is a function of the probability of a collision, and of the consequence of the collision. Thus we can write:

$$Fatality\ Rate = F(Consequences, Probability).$$

Simple physics indicates that the consequence of a collision is a function of crash speed; and simple logic indicates that the probability of collision is a function of the dispersion of speeds on a given highway—more passing means more chances to collide. Thus:

$$\textit{Fatality Rate}$$
$$= F(\textit{Speed}, \textit{Speed Variance}, \textit{Other Factors}).$$

Operationalizing *Speed* as the average speed, and operationalizing *Speed Variance* (85th percentile − average speed) where *FR* denotes *Fatality Rate*:

$$(1)\quad FR = a + b_1 Avg + b_2(85th\% - Avg) + e$$
$$= a + b_2 85th\% + (b_1 - b_2) Avg + e$$
$$(2)\quad = a + b_2 85th\% + b_3 Avg + e.$$

Suppose that speed variance has more effect on fatalities than does speed, per se—"coordinating" the traffic flow is more important than "limiting" it. Then in equation (1), b_2 will be larger than b_1. Since b_3, in equation (2), combines the effect of speed and speed variance, b_3 will actually be negative. That is, in equations of the form of (2), we would expect to get oppositely signed pairs of regression coefficients, b_2 positive and b_3 negative.

Table 3 shows the result of estimating equation (2) on the twelve subsets of the data. (In addition to the speed measures, it also includes a measure of access to emergency medical care,[2] and a measure of driver characteristics.)[3] The results confirm the model: 10 out of 12 of the regression coefficients of *Average Speed* are negative, and 10 out of 12 of the regression coefficients of *85th % Speed* are positive. The coordination effect is larger than the limit effect.

TABLE 3—THE COMBINED EFFECTS OF SPEED AND SPEED VARIANCE[a]

Road Type and Year	*Average Speed*	*85th % Speed*	R^2
Rural Interstate 1981	−.24	.20	.62
	(1.8)	(2.3)	
Rural Arterial 1981	−.75	.58	.25
	(3.7)	(2.8)	
Rural Collector 1981	.01	−.01	.00
	(0.0)	(0.1)	
Urban Freeway 1981	−1.3	.55	.29
	(1.7)	(0.7)	
Urban Interstate 1981	−.04	.10	.12
	(0.3)	(1.2)	
Urban Arterial 1981	−.58	.50	.15
	(2.4)	(2.2)	
Rural Interstate 1982	−.21	.19	.52
	(2.3)	(2.5)	
Rural Arterial 1982	−.41	.35	.08
	(2.0)	(1.8)	
Rural Collector 1982	−.09	.001	.10
	(0.7)	(0.0)	
Urban Freeway 1982	−.39	.30	.14
	(0.8)	(0.7)	
Urban Interstate 1982	.04	−.01	.13
	(0.3)	(0.1)	
Urban Arterial 1982	−.29	.23	.16
	(1.7)	(1.2)	

[a] t-ratios are shown in parentheses; R^2 is corrected for degrees of freedom; and *Hospital Access* and *Driver Characteristics* are also in the equation.

A further interesting result in Table 3 is that the pairs of speed coefficients, in a given regression, tend to be of approximately equal magnitude. But, by definition, $b_3 = b_1 - b_2$, so if b_3 and b_2 are approximately equal, then b_1 must be near to zero. Since, from equation (1), b_1 measures the effect of average speed on the fatality rate, then this would imply that speed, per se, has very little or no effect on fatalities. That is, the limit effect is very small.

[2] We want to measure both the number of hospitals per square mile, and the uniformity of their distribution. The variable used was *Hospitals/Square Mile* multiplied by the *Proportion of Population Living in Nonmetropolitan Areas*. Several variants of this were examined, and this one proved to be superior.

[3] Speeding *Citations per Driver* is a function of both driver aggressiveness and police conscientiousness. The results in Table 4 show a positive coefficient in 10 of the 12 cases, indicating that *Citations per Driver* is primarily a measure of driver behavior. Partial confirmation of this idea is seen in the negative correlation between citations and average driver age: a high proportion of young, presumably aggressive, drivers leads to a high citation rate.

TABLE 4—FINAL REGRESSION EQUATIONS[a]

Road Type	*Speed Variance*	*Citations per Driver*	*Hospital Access*	R^2	N	*Average Speed (if Entered)*[b]
Rural Interstate 1981	.176	.0136	−7.75	.624	41	(−0.5)
	(2.3)	(4.6)	(3.6)			
Rural Interstate 1982	.190	.0071	−5.29	.532	44	(−0.4)
	(2.6)	(2.8)	(3.7)			
Rural Arterial 1981	.677	.0122	.915	.237	46	(−1.3)
	(3.5)	(1.6)	(0.2)			
Rural Arterial 1982	.375	.0116	−.424	.101	47	(−0.5)
	(2.0)	(1.7)	(0.1)			
Rural Collector 1981	.011	.0041	−8.61	.019	41	(−0.1)
	(0.1)	(0.6)	(1.6)			
Rural Collector 1982	.046	.0139	−0.83	.089	41	(−1.2)
	(0.3)	(2.4)	(0.2)			
Urban Freeway 1981	.892	.0634	−.126	.269	19	(−1.2)
	(1.3)	(1.9)	(1.1)			
Urban Freeway 1982	.281	.0410	−2.86	.193	18	(−0.5)
	(0.7)	(2.5)	(0.5)			
Urban Interstate 1981	.103	.0101	.324	.139	26	(0.7)
	(1.2)	(2.0)	(0.2)			
Urban Interstate 1982	−.011	.0106	−.168	.167	27	(0.3)
	(0.2)	(2.8)	(0.1)			
Urban Arterial 1981	.526	−.0187	−1.93	.177	23	(−0.6)
	(2.4)	(1.9)	(0.5)			
Urban Arterial 1982	.304	−.0068	−5.72	.168	21	(−1.0)
	(1.9)	(1.2)	(2.2)			

[a] t-ratios are shown in parentheses; R^2 is corrected for degrees of freedom.
[b] Shows the t-ratio of the *Average Speed* variable if it were to be added to the equation (its potential significance and sign).

A. *A Direct Measure of the Effect of Speed*

We can test this implication by estimating equation (1): it gives direct coefficient estimates for the separate effects of speed and speed variance. When these regressions were run, *Average Speed* was insignificant in all 12 equations, and actually negative in 10 of them. I also tried replacing *Average Speed* with three other speed measures—percent of cars exceeding 55 mph, percent of cars exceeding 65 mph, 85th percentile speed—but results were no different. Once the effect of variance is held constant, there is no discernable effect of speed on the fatality rate.[4]

Table 4 shows the final regression equations. *Average Speed* has been removed from the equations, but the last column (*Average Speed (if Entered)*) indicates the significance and sign it would have if it were to be included—it is not only insignificant but actually has a perverse sign in 10 of the 12 equations. As expected, *Hospital Access* plays its biggest role on rural interstates, the highways that are far removed from normal medical services. And, as expected, the effect of *Speed Variance* is least on relatively uncongested, multilane highways—the rural and urban interstates.

B. *Supporting Evidence*

These results are not unprecedented in the traffic engineering literature, though they do seem to have been forgotten. Twenty years ago, David Solomon investigated the relation

[4] This conclusion is not contradicted by the observed drop in fatalities following the imposition of the 55 mph NMSL in 1973, since speed variance fell that year.

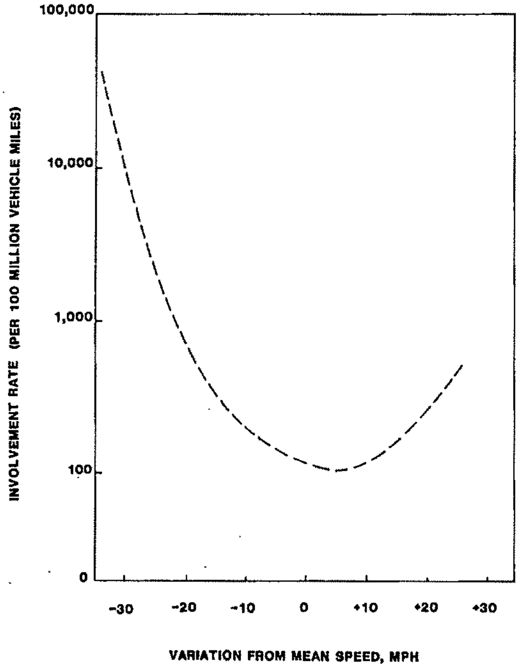


FIGURE 1. DEVIATION FROM AVERAGE SPEED VS. THE COLLISION RATE

between accident rates and variance and plotted the curve in Figure 1. This shows that it is safest to drive at the median speed, and increasingly dangerous to deviate from this speed in either direction; that is, slow drivers are equally responsible for causing accidents. Julie Anna Cirillo (1968) replicated the Solomon curve on interstate highways; and Ezra Hauer (1971) provided a theoretical foundation for the Solomon curve: he derived the number of overtakings expected at various speeds (for example, if I drive at 45 mph, while the median of the pack is 60 mph, how many cars will overtake me per hour, and hence have a chance to collide with me), and showed that his theoretical distribution was nearly identical to the Solomon curve.

III. Discussion

This paper presents evidence that speed laws should be viewed as devices for coordinating speed, not just limiting it. Both the slow driver and the fast one impose negative externalities on the median driver. Apparently, this is a novel conclusion: all current safety campaigns emphasize that "speed kills." They imply that the slower driver is the virtuous one and is helping protect himself and other drivers. It isn't so. To reduce fatalities, it is important that everyone drive at about the same speed. Thus the major consideration in choosing a speed limit is that it be obeyed. And the major consideration for the police is to reduce variance, not speed, because slow drivers are as much a public hazard as fast ones.

Clearly, the 55 mph NMSL ignores these considerations. It focuses on average speed to the exclusion of everything else. Even its compliance mechanism is ill-conceived: any state where more than half the drivers exceed 55 mph is subject to loss of federal highway subsidies. Thus there is as much federal sanction for a 56 mph driver as for a 76 mph driver.[5]

Although I have found no statistically discernible effect from speed, per se, this does not necessarily imply that it is safe to raise the speed limit, for we do not know what effect a higher limit would have on the speed variance. In the twelve data sets examined, there is generally a *negative* correlation between average speed and speed variance (8 negative correlations, 3 positives, and one 0.0); but I take these correlations to be suggestive rather than predictive.

However, the results presented here, and supported by the apparently forgotten observations in the highway engineering literature, do imply that major changes in the National Maximum Speed Limit and police behavior are warranted.

[5] Of course the NMSL was instituted to save energy, not lives, but its energy effects are relatively trivial—approximately a 0.2–1.0 percent reduction in gasoline consumption (Glenn Blomquist, 1984).

REFERENCES

Cirillo, Julie Anna, "Interstate System Accident Research Study II, Interim Report II," *Public Roads*, August 1968, *35*, 71–75.

Blomquist, Glenn, "The 55 mph Speed Limit and Gasoline Consumption," *Resources and Energy*, March 1984, *6*, 21–39.

Forester, Thomas, McNown, Robert and Singell, Larry, "A Cost-Benefit Analysis of the 55 mph Speed Limit," *Southern Economic Journal*, January 1984, *50*, 631–41.

Hauer, Ezra, "Accidents, Overtaking and Speed Control," *Accident Analysis and Prevention*, January 1971, *3*, 1–12.

Jondrow, James, Bowes, Marianne and Levy, Robert, "The Optimal Speed Limit," *Economic Enquiry*, July 1983, *21*, 325–36.

Kamerud, Dana B., "The 55 mph Speed Limit: Costs, Benefits, and Implied Tradeoffs," *Transportation Research*, January 1983, *17A*, 51–64.

Schelling, Thomas C., *Micromotives and Macrobehavior*, New York: W. W. Norton, 1978.

Solomon, David, "Accidents on Main Rural Highways Related to Speed, Driver, and Vehicle," Federal Highway Administration, U.S. Department of Transportation, July 1964.

Federal Highway Administration, *Highway Statistics*, U.S. Department of Transportation, Washington, various yearly issues.

Identification by Disaggregation

By Matthew J. Cushing and Mary G. McGarvey*

Standard economic theory predicts that the actions of individual participants in competitive markets have negligible effects on market-determined aggregates. Applied researchers,[1] and even some econometric textbooks,[2] incorrectly infer from this that market prices can be modeled as econometrically exogenous with respect to the quantity demanded of an individual consumer.[3] This faulty inference has even led some researchers (for example, Robert Engle, 1978; Nicholas Kiefer, 1984; Roger Waud, 1974) to employ an estimation strategy we call identification by disaggregation (*IBD*). This procedure attempts to circumvent the simultaneity problem in a macro regression by disaggregating the dependent variable and estimating the relationship for individual agents or sectors. This note provides a simple proof that estimates using disaggregated dependent variables suffer, on average, from the same degree of simultaneity bias as the estimates using aggregate data.

Let Y be a $T \times 1$ vector of T observations on a macro variable and X be a $T \times k$ matrix of T observations on k macro variables. For example, Y and X could be aggregate output growth and money growth. Let the aggregate relationship between Y and X be

$$(1) \qquad Y = XB + e,$$

where all variables are deviations from their means and $\text{plim}(X'X/T)^{-1} = Q$ and $E(e) = 0$. Assume that there is some feedback from X to Y so that $\text{plim}(X'e/T) = P \neq 0$. The ordinary least square (*OLS*) estimation of B from (1) yields the inconsistent estimator $\hat{B} = (X'X)^{-1}X'Y$, where the inconsistency is

$$(2) \qquad \text{plim}(\hat{B} - B) = QP.$$

It is often argued that the inconsistency in sectoral regressions is smaller. For example, Waud argues that since the feedback from employment in a particular industry to aggregate money growth should be minimal, the simultaneous equations bias in sectoral regressions should be smaller than that in the aggregate regression.[4]

From (1) we can write the disaggregated relationships between the Y_i and X as

$$(1') \qquad Y_i = XB_i + e_i$$

where $Y = \sum_{i=1}^{n} Y_i/n, \quad B = \sum_{i=1}^{n} B_i/n,$

$$e = \sum_{i=1}^{n} e_i/n$$

and n is the number of sectors considered. The *OLS* estimator of B_i from (1′) is $\hat{B}_i = (X'X)^{-1}X'Y_i$. It is immediately obvious that

*Departments of Economics, Emory University, Atlanta, GA 30322, and Georgia State University, Atlanta, GA 30303, respectively. This note has benefited from helpful comments from Mark Meador and an anonymous referee.

[1] Nicholas Kiefer uses household data to estimate the Rotterdam demand model, arguing "...according to the usual arguments the simultaneity problem is not present. Surely supply to an individual is perfectly elastic..." (1984, p. 288).

[2] "While studying the demand for gasoline by households, we can treat the quantity demanded as endogenous and income and price as exogenous, arguing that the household does not have control over these" G. S. Maddala (1977, p. 5).

[3] The premise is that individual agents or sectors do not control the aggregate variables. However, econometric exogeneity can fail if the error term of the individual behavioral equation is merely correlated with the aggregate variables. For a careful discussion of alternative definitions of econometric exogeneity and their usefulness for inference, see Robert Engle, David Hendry, and Jean-Francois Richard (1983).

[4] However, Waud admits "...it is very difficult to establish unequivocally that the reduced form approach used here definitely reduces single-equation least squares bias, even though it seems likely that it does" (p. 186).

the *OLS* estimator $\hat{B}$ from (1) is simply the average of the *OLS* estimators $\hat{B}_i$ from (1′). Defining $\text{plim}(X'e_i/T) = P_i$, the inconsistency of any $\hat{B}_i$ is

$$(2') \qquad \text{plim}(\hat{B}_i - B_i) = QP_i.$$

Clearly, the inconsistency of the aggregate estimator, (2), is simply the average inconsistency of the sectoral estimators, (2′), over the n sectors. Of course, estimates from sectors with lower than average P_i's exhibit smaller inconsistencies than the aggregate estimator. However, some prior information is required in order to identify these sectors. If sectors are chosen at random, the expected inconsistency of the sectoral estimator is identical to the inconsistency of the aggregate estimator.

The above result is a consequence of the linearity of ordinary least squares. The result holds for any linear estimator (for example, *OLS*, *GLS* with a known covariance matrix, and estimation under linear restrictions). Let $\hat{B} = A'Y$, where A is some $T \times k$ linear transformation matrix. The corresponding sectoral estimator is $\hat{B}_i = A'Y_i$. It is clear that $\hat{B}$ is simply the average of the $\hat{B}_i$'s so that the inconsistency of $\hat{B}$ is the average of the $\hat{B}_i$'s inconsistencies.[5]

The results of this paper allow us to reevaluate the discussion of identification by disaggregation contained in Thomas Cooley and Stephen LeRoy (1981). By working through a specific example we can see how the intuition behind *IBD* fails.

Consider a simple, aggregate money demand function (in deviations from means),

$$(3) \qquad m = ar + e,$$

where m is defined as average money balances, $m = \sum_{i=1}^{n} m_i/n$, and n is the number of sectors. Let the money supply function follow a feedback rule of the form

$$(4) \qquad r = bm + v.$$

[5]Although we consider only linear estimation techniques, there is no reason to believe that disaggregating a nonlinear macro relationship and estimating the sectoral relationships would reduce the simultaneity bias.

To avoid the obvious simultaneity bias in estimating a with aggregate data, Cooley and LeRoy explore the possibility of estimating (3) using sectoral money stock data. The demand for money in the ith sector can be modeled as

$$(3') \qquad m_i = a_i r + e_i.$$

Assume that the sectoral shock e_i is the sum of two mutually uncorrelated components: w, a shock common to all sectors with variance σ_w^2, and u_i, a shock specific to sector i and uncorrelated across sectors with variance σ_i^2. The inconsistency of the sectoral *OLS* estimator is (Cooley-LeRoy, p. 840)

$$(5') \quad \text{plim}(\hat{a}_i - a_i)$$
$$= \text{plim}(r'e_i/T)/\text{plim}(r'r/T)$$
$$= b\left(\sigma_w^2 + \sigma_i^2/n\right)(1 - ba)/\left(b^2\sigma_e^2 + \sigma_v^2\right).$$

Cooley and LeRoy observe that the inconsistency of the sectoral estimator will be small if there are a large number of sectors (n is large) and the error terms (e_i) are mutually uncorrelated (i.e., the common factor is zero). Cooley and LeRoy argue that, in this particular case, *IBD* will not resolve the identification problem because money stock data is available for only four sectors and the assumption of a zero common component is implausible. However, their discussion leaves the impression that, although *IBD* is not useful in this case, it may be useful in other circumstances.

The results of this paper provide a much stronger and more general criticism of this attempt at identification by disaggregation. It is true that if the common shock is zero ($\sigma_w^2 = 0$) and n, the number of sectors, is large (holding σ_i^2 constant), the inconsistency of the sectoral estimator will be small. However, consider the inconsistency of the aggregate estimator

$$(5) \quad \text{plim}(\hat{a} - a)$$
$$= \text{plim}(r'e/T)/\text{plim}(r'r/T)$$
$$= (1 - ba)b\sigma_e^2/\left(b^2\sigma_e^2 + \sigma_v^w\right).$$

Recall that the aggregate shock e is the average of the sectoral shocks. If the number of sectors is large and the shocks are independent, the variance of the aggregate shock must be small. As such, the inconsistency of the aggregate estimator, (5), is clearly small. The conditions under which the inconsistencies of the sectoral estimators are small guarantee that the inconsistency of the aggregate estimator is small. Identification by disaggregation is a strategy that succeeds only when it is unnecessary.

REFERENCES

Cooley, Thomas F. and LeRoy, Stephen F., "Identification and Estimation of Money Demand," *American Economic Review*, December 1981, *71*, 825–44.

Engle, Robert F., "Testing Price Equations for Stability Across Spectral Frequency Bands," *Econometrica*, July 1978, *46*, 869–81.

______, Hendry, David F. and Richard, Jean-Francois, "Exogeneity," *Econometrica*, March 1983, *51*, 277–304.

Kiefer, Nicholas M., "Microeconometric Evidence on the Neoclassical Model of Demand," *Journal of Econometrics*, July 1984, *25*, 285–302.

Maddala, G. S., *Econometrics*, New York: McGraw-Hill, 1977.

Waud, Roger N., "Monetary and Fiscal Effects on Economic Activity: A Reduced Form Examination of their Relative Importance," *Review of Economics and Statistics*, May 1974, *56*, 177–87.

Race, Recent Entry, and Labor Market Participation

By EUGENE SILBERBERG*

Economists and other social scientists have long noted differences in incomes and general labor market success between ethnic groups. In recent years particular attention has been focused on the experiences of black vis-à-vis white workers, with particular reference to income differences that might be attributable to racial discrimination. The model is typically specified with a linear regression equation,

$$(1) \qquad y = \beta X + \gamma D + u$$

when y is some measure of labor market success, X is a vector of explanatory variables related to productivity, with a random disturbance term u, and where D is a dummy variable for race (for example, $D=1$ for black workers). A significant negative value of γ is taken to represent the possible existence of labor market discrimination. Accumulated formal education, reported in census data, is frequently used as an explanatory variable in these analyses even though that variable is understood to be a poor proxy for productivity.[1]

It is well-known that if the mean y observation is less for blacks than for whites, spurious confirmations of discrimination can occur if additional variables affecting productivity are omitted from the regression, or if the X variables contain measurement errors.[2] This paper explores two aspects of this issue: 1) there is a theoretical reason to believe that new entrants to a higher-paying occupation will exercise their earning capacities to a lesser extent than workers already familiar with the field; that is, the labor supply curve of new entrants is apt to be to the left of the supply curve of other workers; and 2) data and empirical analyses are presented that confirm this theory and that cast serious concern over the adequacy of various census-type data in investigations of labor market discrimination.

I. An Economic Theory of Differential Labor Supply

A common stereotype is the observation by "people of wealth" that the *nouveau riche* tend to spend their money in ways which are considered "gauche" by the "old rich." I would argue that this phenomenon is real, and based on positive information costs. The standard utility-maximum model not only takes prices as exogenously determined (an assumption maintained here), but also the set of goods over which the consumer chooses. In fact, the goods and services individuals decide to *consider* purchasing depends on their previous experiences in the consumer goods market. People of very limited means will not have had much exposure to a wide variety of goods regularly consumed by people whose incomes are greater. For example, poor people on average will have scant experience, *and therefore scant information* about the pleasures of skiing, gourmet restaurants, tailor-made clothes, and the like.

Following George Stigler (1961), the information consumers possess about goods is not free. It is generally acquired, in the case of

*Professor of Economics, University of Washington, Seattle, WA 98195. I am indebted to Yoram Barzel, Masanori Hashimoto, Levis Kochin, and an anonymous referee for their valuable comments, and to Joyce Hershberger for her insights and skill in processing the data. Any remaining errors are my own.

[1] See, for example, Gary Becker (1957), Richard Freeman (1978), James Gwartney (1970), William Landes (1968), Finis Welch (1967, 1973). See also Leonard Weiss and Jeffrey Williamson (1972) and Christopher Jenks et al. (1972) on the effects of environmental variables on labor market success.

[2] See Masanori Hashimoto and Levis Kochin (1980).

frequent and repetitively consumed goods, through actual purchase and experience. For less-frequently purchased goods such as automobiles, stereo or photographic equipment, etc., some time and resources will likely be invested. As a result, the *number* of goods consumed, as well as the levels of each good, will in general differ for two individuals, even if they have identical tastes and identical incomes, if one of those individuals has "always" had the same income and the other person has only recently acquired that income. These two people will have different information on what to consume.

The different consumption patterns of the above two hypothetical people are to some extent predictable. We should expect the person with newly acquired larger income to consume fewer commodities than the other person. Correspondingly, the observed income elasticities of the goods previously consumed should be greater, the more recent the larger income became available.[3] In particular, one of the goods consumed by all individuals to one extent or another is leisure; the above analysis suggests that *people whose wage rates have risen suddenly will not, on average, exercise that higher earning capacity to the extent done so by people who are more accustomed to that higher wage*. People who are new to a higher-wage situation can be expected, on average, to consume more leisure than workers who have earned the higher wage for a longer period of time. As a person begins acquiring information about new goods due to a higher income, the number of goods consumed will expand. The effect on the consumption of the previously consumed goods will depend on whether those goods are substitutes or complements to the new good or goods.[4] Similar conclusions follow for unexpected decreases in income. The *nouveau pauvre* will initially consume a greater number of goods, with generally smaller income elasticities, than in the longer run. Moreover, they will be less "efficient" at coping with their new income than those accustomed to such incomes.[5]

More formally, consider the compensated demand for some good X_i, denoted $X_i(P,U)$, where $P=(P_1,\dots,P_{n+1})$ is an $n+1$ dimensional price vector, $X_1,\dots,X_n$ represent the original goods consumed, X_{n+1} represents a composite commodity that is consumed after the consumer acquires information about its availability, price, etc., and U is the initial utility level. The initial situation is where $X_{n+1}=0$; X_{n+1} is "fixed" at zero in the sense that this good is not at first even considered as a candidate for consumption by the consumer. Define the "conditional" compensated demand function for X_i when X_{n+1} is fixed as $X_i^{n+1}(P, X_{n+1}, U)$, where the argument X_{n+1} in this demand function is meant to be a constant value.[6] The fundamental relationship between the demand for X_i when X_{n+1} is variable as opposed to when it is fixed is

$$(2)\quad X_i(P,U)=X_i^{n+1}(P,X_{n+1}(P,U),U).$$

Differentiation of (2) with respect to P_i and the utility level U yields, respectively,

$$(3)\quad \partial X_i/\partial P_i=\partial X_i^{n+1}/\partial P_i + (\partial X_i^{n+1}/\partial X_{n+1})(\partial X_{n+1}/\partial P_i) \qquad i=1,\dots,n;$$

[3]To take an extreme example, consider the goods that might be consumed by children, if they suddenly received adult incomes. They would probably exhibit greater income elasticities of video games and hamburgers than most adults.

[4]From the consumer's budget equation, the income elasticities weighted by the shares in consumption sum to unity: $\Sigma_i\eta_i=1$, where $\Sigma_i=P_iX_i/I$ and $\eta_i=(I/X_i)(\partial X_i/\partial I)$. It follows from this simple relationship that if an additional good X_{n+1} is consumed, the total share in the budget of the previous n commodities must have fallen. For this to happen, some goods must have become more income inelastic than previously. Although the shares of income spent on some particular good or goods may increase, the weighted income elasticities must decrease.

[5]Talmudic law in fact considers as worthy recipients for charity people who have suffered drastic decreases in income, even if their incomes remain high relative to their community.

[6]The concept of conditional demands was originated by Robert Pollak (1969).

$$(4)\quad \partial X_i/\partial U = \partial X_i^{n+1}/\partial U + (\partial X_i^{n+1}/\partial X_{n+1})(\partial X_{n+1}/\partial U) \qquad i=1,\dots,n.$$

The standard LeChatelier results for these compensated demand functions (see my 1978 study, ch. 9), together with (3), yield

$$(5)\quad (\partial X_i^{n+1}/\partial X_{n+1})(\partial X_{n+1}/\partial P_i) = \partial X_i/\partial P_i - \partial X_i^{n+1}/\partial P_i \le 0 \qquad i=1,\dots,n.$$

The two terms on the left must be of opposite sign. The sign of the term $\partial X_{n+1}/\partial P_i$ (that is identically $\partial X_i/\partial P_{n+1}$ for these compensated "cross effects") defines whether the composite good X_{n+1} is a net substitute or complement to X_i. If X_i and X_{n+1} are net substitutes ($\partial X_i/\partial P_{n+1} > 0$), then $\partial X_i^{n+1}/\partial X_{n+1} \le 0$, and $\partial X_i^{n+1}/\partial X_{n+1} \ge 0$ if X_i and X_{n+1} are net complements.

Turning now to equation (4), the term $\partial X_{n+1}/\partial U$ must initially be positive, since initially $X_{n+1}=0$; X_{n+1} becomes positive as income (and utility) increases. Therefore, $\partial X_i/\partial U \ge \partial X_i^{n+1}/\partial U$ when X_i and X_{n+1} are net complements and $\partial X_i/\partial U \le \partial X_i^{n+1}/\partial U$ when X_i and X_{n+1} are net substitutes. In words, assuming no asymmetric income effects so that the uncompensated demands can be considered, if new information leads to the consumption of some composite good outside the original consumption set, then the income elasticities of goods which are substitutes in consumption to the composite good will fall, whereas the income elasticities of complements to the composite good will rise.

Consider the specific good "leisure" and its response to a change in income in the above context. Consider a worker whose prior wage was, say, \$5 per hour and who earned or expected to earn \$10,000 per year. Suppose also that this worker lived in a neighborhood where \$10,000 was a common annual income, and this person's associates were earning similar amounts. Then this person would have perhaps good information on the goods that might be consumed with \$10,000 annual income, but less or no information on some other goods that might be consumed with, say, \$20,000 annual income.

Suppose now that this person's wage doubles to \$10 per hour. Then the above analysis suggests that the *initial* consumption of leisure will be higher than after the passage of some time.[7] An increase in the wage rate of course always, by raising the opportunity cost of leisure, produces a substitution effect away from leisure. This is offset to some extent by a wealth effect, leisure being a normal good. However, we should expect the wealth effect on leisure to be greater in the period immediately following the wage increase than after, say, a few years, when this person will have acquired more information on additional goods to consume. It is possible that some of these goods will be complementary to leisure, for example, a large screen television. Because of this, a refutable hypothesis logically implied by the utility-maximization model is not possible.[8] However, given the net tendency towards substitutes first outlined by John Hicks in *Value and Capital* (1937), we should expect workers with a recently acquired higher wage to have a larger income elasticity of leisure than comparable workers with a longer history of the higher wage.[9]

[7]For the same kinds of reasons, Armen Alchian (1959) proposed that costs would increase if production took place at a more rapid rate, holding the total volume of production constant, and that planning further into the future would be cost saving. Thus the temporarily high consumption of leisure as wage rates rise can be regarded as movement along a least-cost path of dynamic adjustment to the higher income levels.

[8]In a two-good world consisting of the previously consumed goods as one composite and the new goods as a second composite commodity, net substitution and hence larger income effects on the old good is implied.

[9]Reuben Gronau (1977) correctly distinguishes leisure from "work at home." An increase in the market wage rate leads to substitution towards labor market work from work at home, whereas the effect on leisure is the indeterminate result of an income effect and reduction of total time spent at home. Since the effects of work at home would already be well known, one would again expect less substitution of market work for work at home, when the wage rate rises, for workers for whom the higher market rate is "new." This would appear in the data as less time spent in labor market activity for

Observations that might appear to contradict this theory are the "upwardly mobile" individuals who vigorously exercise their earning capacity in order to rapidly increase their income. These people, however, seem to be acting as if they had very good information about the benefits of higher income. The above analysis is most applicable when the increase in earning capacity is *unanticipated.* If a person has been planning for this higher income, then when the higher earning capacity is realized, that person's consumption will more closely resemble those whose incomes have been at that higher level for some time, because more information will have been procured regarding the possibilities for consumption with that larger income.[10]

II. Empirical Analysis

This theory will now be tested with data relating to a group of black workers for whom a newly acquired higher wage could reasonably be considered a windfall (but permanent) gain. After being excluded from various construction unions, these workers were suddenly admitted to several unions in the Seattle area as a result of a court order. This group of black workers were new entrants to much higher-paying jobs than they were previously capable of obtaining. The nonblack workers, too, were obtaining training and relatively high-wage jobs. However, the white workers were much more likely to be familiar with these crafts and the incomes that could be earned than that first cohort of black workers that was admitted to the union under the court order. It was common for the white workers to have relatives and friends in the union; this was in fact of some importance in their gaining entrance into the union. There was no chance, however, that the black workers had relatives or black acquaintances in these unions earning the high (relative to average nonagricultural wages) journeyman rates. It is therefore plausible that in the total cohort of workers indentured into these craft unions in the mid-1970's, the black workers anticipated lower permanent incomes than the white workers, and were less apt to have planned for the higher incomes that could be achieved.[11]

Data about black and nonblack[12] workers indentured into the plumbing and pipefitting trades (one union, but two distinct crafts) and the electricians' union during the 1970's were obtained from the unions, employer groups, and apprenticeship schools. Wages are uniform in each craft, but income differences due to hours worked were apparent. Black workers worked fewer hours per month, on average, than their nonblack counterparts; the differences were statistically significant at the .01 level of significance. Observations on hours worked were taken over the 18-month period April 1979 through September 1980.

such workers. This observation is confirmed by the data in this paper, but the allocation of nonmarket time to leisure vs. work at home is not analyzed. In this paper, leisure means "nonmarket" activity.

[10]The above reasoning bears a superficial resemblance to that used by Milton Friedman (1957) in the permanent income hypothesis. The permanent income hypothesis asserts that people with temporarily high income will engage in higher savings than those with permanent income of the same amount, in order to even out their flow of consumption over time. The analysis in this paper, however, assumes that the earning capacities of the two individuals are *both* permanently higher. The difference in behavior results from the differential information possessed by individuals whose higher wages are "new," i.e., not fully anticipated, vs. those who were familiar with the higher wages. Thus the analogy with the permanent income hypothesis is somewhat faulty. An interesting possibility, however, is that individuals faced with an unanticipated permanent wage increase will nonetheless anticipate that their information about goods will increase in the future, leading to a substitution of saving for leisure. This would to some extent, at least, offset the larger income elasticity of leisure for these individuals.

[11]Some of the rents received through membership in these unions would have been already dissipated by the resources spent to gain entrance. However, given the large role played by third parties (i.e., the federal government), it seems most plausible that union membership conferred real gains to the new entrants. Thus the observed behavior is not attributable to pure substitution effects.

[12]The workers were classified as "black" and nonblack in the original data. There were only a handful of nonwhites, mainly people of Asian descent, in the category "nonblacks."

TABLE 1—ELECTRICIANS: MEANS AND STANDARD DEVIATIONS

Apprentices	Blacks (22)	Nonblacks (259)
Hours Worked (*HOURS*)	146.49	153.18
	(12.92)	(14.17)
Month/Year Born (*DOB*)	52.87	54.18
	(3.56)	(3.22)
Apprentice School *GPA* (*APPGPA*)	79.73	88.29
	(5.11)	(4.84)
High School *GPA* (*GPA*)	2.14	2.64
	(0.53)	(0.56)
Adverse Comments (*COMMENT*)	0.59	0.26
	(0.50)	(0.44)
Years of Education (*YRSEDUC*)	12.76	13.14
	(1.06)	(1.32)
Journeymen	Blacks (48)	Nonblacks (64)
Hours Worked (*HOURS*)	138.93	148.58
	(24.55)	(19.77)
Quits/Discharges (*QUIT/DIS*)	2.14	0.52
	(2.68)	(0.91)
Time Between Jobs (*VOLTWJOB*)	11.10	3.55
	(20.03)	(12.30)
Jobs Turned Down (*TURNDOWN*)	4.35	1.61
	(7.39)	(5.91)
Voluntary Time Off (*VOLTIMEO*)	0.45	0.27
	(0.42)	(0.41)

Notes: COMMENT: Dummy variable indicating that an apprentice had received an unfavorable evaluation by job foreman or supervisor, 18-month period; *QUIT/DIS*: Number of times in 18-month period a journeyman quit or was fired from job; *VOLTWJOB*: Interval (days) between termination of one job and sign-up at union hall for new job; *TURNDOWN*: Number of jobs voluntarily refused by journeyman; *VOLTIMEO*: Hours lost due to leaving job site early.

A. *Electricians*

The data available for the electricians are shown in Table 1. Differences in the means of these variables, for black and nonblack workers, are apparent by inspection; they are all in fact different at the 5 percent level of significance.

A regression using just the data on formal education produced (in all regression equations, y denotes hours worked per month; $D=1$ for blacks)

$$\begin{aligned} y = 153.17 &+ 2.89GPA - 0.59YRSEDUC \\ &\;\;(1.80) \qquad (-0.862) \\ &-5.23D \\ &\;(1.615) \qquad\qquad \bar{R}^2 = .04 \end{aligned}$$

Ignoring the insignificant years of education variable, grade point average (GPA) did show a positive and significant coefficient, but only 1.5 hours per month were attributable to this variable. The largest "explanation" was the dummy variable for race, significant at just over the 5 percent level, and accounting for 5.23 hours per month difference in hours worked between blacks and nonblacks. Taken at face value, the above regression would confirm the presence of racial discrimination for these workers. However, the question remains: do data on formal education, GPA, and length of schooling adequately measure productivity, or are the results simply due to the use of poor proxies?

The importance of the race variable changed dramatically when variables more directly related than formal education to job training and performance were used in the regressions. Using these variables, typical re-

sults were:

Apprentices:

$$y = 116.72 \underset{(5.86)}{} - \underset{(.02)}{.05RACE} - \underset{(-1.35)}{.34DOB}$$

$$+ \underset{(3.92)}{.66APPGPA} - \underset{(-2.42)}{4.41COMMENTS}$$

$$\bar{R}^2 = .08$$

Journeymen:

$$y = \underset{(61.08)}{152.03} + \underset{(.01)}{.04RACE} - \underset{(-4.42)}{4.41QUITS/DIS}$$

$$- \underset{(2.81)}{.33VOLTWJOB} \qquad \bar{R}^2 = .25$$

For the apprentice electricians, the race variable became insignificant ($\beta = -1.06$, $t = -.31$) just by substituting the apprenticeship grades (*APPGPA*) for high school grades (*GPA*). With this one variable, 84 percent of the difference in hours worked by blacks vs. nonblacks was explained. This would seem to confirm the hypothesis that bad measurement of potential productivity rather than discrimination is the source of some of the observed differences by race in labor market success. The most dramatic variable in the raw data for the journeyman electricians was the observed difference in the interval between termination from one job and signing up for work at the union hall (*VOLTWJOB*): 11.10 days for blacks vs. 3.55 days for nonblacks; this variable accounted for a difference of 2.49 hours worked per month. The combined effects of quitting and being discharged (almost exclusively for tardiness and absence) had the largest impact. These variables seem to reflect different supply responses on the part of those two groups of workers.

B. *Plumbers and Pipefitters*

A more elaborate set of data were available for the plumbing and pipefitting trades.[13] Data on the plumbers and pipefitters came from two sources, the Joint Apprenticeship and Training Council (JATC) which runs the apprenticeship training program, and the union itself. During apprenticeship, workers must attend two three-hour evening classes per week for four years. The apprentices encounter about two dozen separate instructors, whose salaries are paid by the state. These school records were obtained for each worker, journeyman as well as for individuals who were apprentices during the sample period.

From these records the average number of absences from apprenticeship school per year and a grade point average were obtained. Also, during the apprenticeship period, foremen or supervisors on the job make regular evaluations of each worker, and rate the employees as poor, fair, good, very good, or excellent. By scaling the latter categories from zero to 4, an "on-the-job *GPA*" (*OTJGPA*) was constructed for each individual. In this sample, approximately 250 separate foremen and supervisors were involved with these evaluations. Additionally, the number of times the employer indicated that an individual was "terminated for cause," that is, terminated for reasons other than reduction in force due to completion of a job, was noted. The "causes," as in the electrical trade, almost invariably were absences and tardiness.

Additional variables were obtained from the union. Jobs are dispatched from this local by telephone. When a job comes in (usually early in the morning) the dispatcher starts at the top of the out-of-work list and telephones each worker in the order they appear on the list. Notation is made if there is no response to the call, or if the job is turned down. A worker may freely turn down two jobs, but after three turndowns, he or she moves to the bottom of the list. Also noted on these work cards are the times an

[13]Although these two trades are organized in one union in Seattle, they are distinct crafts, with separate dispatch procedures. Plumbers generally work in small groups (two or three is typical even on large commercial jobs) whereas pipefitters (also called "steamfitters" by some employers) work in large groups often numbering in the hundreds, such as on nuclear power plants.

TABLE 2—PLUMBERS AND PIPEFITTERS: MEANS AND STANDARD DEVIATIONS

	Pipefitter Apprentices		Pipefitter Journeymen	
	Blacks (12)	Whites (31)	Blacks (26)	Whites (12)
Hours Worked	135.54	154.14	133.78	151.18
(*HOURS*)	(29.11)	(16.82)	(26.96)	(12.57)
Class Absences	4.46	2.84	6.40	3.82
(*ABSENCE*)	(3.66)	(5.84)	(5.23)	(2.32)
Class *GPA* (*GPA*)	2.48	3.13	2.35	2.95
	(0.62)	(0.49)	(0.59)	(0.45)
On-the-Job *GPA*	2.46	3.09	2.69	2.88
(*OTJGPA*)	(0.56)	(0.43)	(0.67)	(0.38)
Quits/No Rehire	0.17	0.10	0.58	0.17
(*QUITNOREHIRE*)	(0.58)	(0.54)	(1.06)	(0.58)
Area Preference	0.17	0.03	0.38	0.33
(*AREACODE*)	(0.39)	(0.18)	(0.50)	(0.49)
Termination for	1.92	0.26	1.08	0.67
Cause (*TERMCAUSE*)	(3.34)	(0.58)	(2.37)	(1.30)
Nonresponse to Job	0.58	0.32	5.04	1.92
Call (*NORESP*)	(1.16)	(1.47)	(8.20)	(3.09)
	Plumber Apprentices		**Plumber Journeymen**	
	Blacks (19)	Whites (46)	Blacks (18)	Whites (9)
Hours Worked	137.89	151.90	128.46	153.08
(*HOURS*)	(22.49)	(11.54)	(27.37)	(17.37)
Class Absences	5.69	2.06	5.72	3.42
(*ABSENCE*)	(5.17)	(1.71)	(4.72)	(2.64)
Class *GPA* (*GPA*)	2.00	2.78	2.08	3.20
	(0.56)	(0.62)	(0.56)	(0.45)
On-the-Job GPA	2.53	2.96	2.60	3.29
(*OTJGPA*)	(0.42)	(0.62)	(0.53)	(0.35)
Quits/No Rehire	0.21	0.0	0.78	0.22
(*QUITNOREHIRE*)	(.535)	(0.0)	(1.40)	(0.44)
Area Preference	0.21	0.87	0.50	0.33
(*AREACODE*)	(0.42)	(0.35)	(0.51)	(0.50)
Termination for	0.84	0.30	0.78	0.0
Cause (*TERMCAUSE*)	(1.50)	(0.87)	(1.52)	(0.0)
Nonresponse to Job	0.68	0.0	2.22	0.56
Call (*NORESP*)	(1.20)	(0.0)	(3.41)	(1.67)

Notes: Hours Worked (*HOURS*) = average number of hours worked per month, 18-month period; Class Absences (*ABSENCE*) = average days per year absent from apprenticeship classes; On-the-Job *GPA* (*OTJGPA*) = average grades received from supervisors during apprenticeship on-the-job training; Quits/No Rehire (*QUITNOREHIRE*) = number of times voluntarily quit a job and/or number of "not eligible for rehire" entries upon termination from a job, 18-month period; Area Preference (*AREACODE*) = zero, will work anywhere; one, work only in King County; two, work only in eastern Washington; Termination for Cause (*TERMCAUSE*) = number of times terminated from job for reason other than rotation or end of job during apprenticeship training; Nonresponse to Job Call (*NORESP*) = number of times did not respond to a call for dispatch to a job, 18-month period; Class *GPA* (*GPA*) = average grade point received from apprenticeship classes completed.

employee quit a job or was listed by the employer as "not eligible for rehire." Additionally, an area preference code was noted. Workers are allowed, without penalty, to indicate a preference for not traveling out of the immediate metropolitan area for a job.

The mean values and standard deviations (in parentheses) for each of these variables for plumbers and pipefitters, apprentices and journeymen are displayed in Table 2 for black and nonblack workers. It is apparent by inspection that the black and nonblack

TABLE 3—DEPENDENT VARIABLE: HOURS PER MONTH

Employee Group	Constant	*RACE*	*ABSENCES*	*OTJGPA*	*NONRESPONSE*	R^2
Apprentice Plumbers	131.58	.082	−2.833	8.816		.583
	(16.76)	(.024)	(−6.264)	(3.543)		
Journeyman Plumbers	76.43	−3.947	−1.571	24.931		.470
	(2.40)	(−.405)	(−1.514)	(2.776)		
Apprentice Pipefitters	100.18	−5.443	−.970	18.377		.299
	(5.11)	(−.727)	(−1.773)	(2.953)		
Journeyman Pipefitters	134.62	−11.106	−1.861	8.237		.249
	(6.63)	(−1.438)	(−2.244)	(1.310)		
Apprentice Plumbers	130.60	.932	−2.562	8.959	−2.595	.584
	(16.54)	(.269)	(−4.947)	(3.600)	(−1.072)	
Journeyman Plumbers	88.22	−2.573	−1.217	21.444	−2.744	.539
	(2.92)	(−.283)	(−1.238)	(2.510)	(−2.103)	
Apprentice Pipefitters	105.12	−6.033	−.889	16.891	−1.829	.293
	(5.11)	(−.799)	(−1.594)	(2.598)	(−.825)	
Journeyman Pipefitters	146.91	−6.009	−1.249	4.711	−2.343	.688
	(11.12)	(−1.195)	(−2.307)	(1.154)	(−6.992)	

Note: The *t*-statistics are shown in parentheses.

workers differed in these attributes. In general, the nonblack workers performed at higher levels in the apprenticeship program, with respect to both skill acquisition and motivational factors such as class attendance and absences on the job; however, not all of these differences were statistically significant in pairwise comparisons by race.[14] Employer evaluations were on average higher for the nonblacks though these means obscure some very high marks given to many black workers.

When these variables were used, the differences in hours worked per month that were attributable to the dummy variable for race (*RACE*) became insignificant. Typical regressions are presented in Table 3.[15] Absences (from apprentice school) alone explained all significant differences in hours worked per month between the black and nonblack plumber apprentices, and *RACE* became insignificant at 5 percent for all four groups with the addition of apprentice school grades (*GPA*). The inclusion of *OTJGPA* improved the regressions further. The *NORESPONSE* variable was particularly significant in the regression for the journeyman pipefitters—the simple correlation between this variable and hours worked was .809 for the *black* journeyman pipefitters.[16]

[14] The variables for which the means by race are significantly different at the 5 percent level are: 1) Pipefitter Apprentices: On-the-Job *GPA*, Quits/No Rehire, Termination for Cause, and (Apprentice) Class *GPA*; 2) Pipefitter Journeymen: Class Absences, No Response to Job Call, Class *GPA*; Quit/No Rehire Significant and just over 5 percent; 3) Plumber Apprentices: All variables except Termination for Cause, which was significant at 10 percent; 4) Plumber Journeymen: On-the-Job *GPA*, Class *GPA*; Class Absences, Termination for Cause, and No Response to Job Call significant at 10 percent.

[15] The variables in Table 2 were correlated with each other in predictable ways, producing some minor multicollinearity problems in the regression results. In general, apprenticeship school grades (*GPA*) were correlated with On-the-Job *GPA* (*OTJGPA*), and both were negatively correlated with absences in apprenticeship school. These correlations were in general significant at the 5 percent level. Terminations for cause were similarly correlated in the obvious ways with apprentice school performance.

[16] "No responses" accounts for a difference of 7.31 hours worked between black and nonblack journeyman pipefitters. One might be tempted to question whether these dispatch phone calls were done without regard to race. In fact it would be difficult to discriminate in that fashion over an 18-month period, the time of this sample. The union keeps the work cards for public viewing by the journeymen; such a policy would be easily detected by observing the places on the list from which the workers were dispatched.

III. Conclusions

The data and regressions for the workers in this sample show behavioral differences not easily explainable on the basis of census-type data. Sociologists use the phrase "socialization" to refer to the process of adapting to the norms of some environment. The theory presented here perhaps explains part of this socialization process. It shows why relatively more leisure will be consumed by workers new to a higher wage environment: in this case, workers who for the first time were admitted to craft unions. The adjustment process described earlier explains in a refutable way the appearance in the data of differences in behavior such as in attendance in apprentice school, in the time to sign the "out-of-work" list for new jobs, or, on a more elementary level, in quits and discharges for tardiness or absences. These actions, all apparently voluntary on the part of individuals, may, in other samples, explain part or all of the observed differences by race in labor market performance.

It is possible, of course, that any particular observed difference between black and non-black workers, for example, differences in education reported by the U.S. Census Bureau, or in some of the variables noted above, are due to either real or perceived discriminatory treatment of blacks. That is, poorer evaluations in school or on the job can be attributed to a hostile environment for blacks that lowered the real wage to those workers. However, it is difficult to imagine an experiment that would render such a hypothesis falsifiable. Absent specific evidence, and given the number of individuals involved in the various evaluations used in the data in this paper, this interpretation has been heavily discounted, since all observed differences by race would thereby automatically be evidence of discrimination. Given that economic theory suggests that, under the test conditions apparent in this experiment, such behavioral differences are to be expected, it is more reasonable to regard these results as confirmation of that theory rather than of capricious (and pervasive) discrimination.

REFERENCES

Alchian, Armen A., "Costs and Outputs," in M. Abramovitz et al. eds., *The Allocation of Economic Resources: Essays in Honor of Bernard F. Haley*, Stanford: Stanford University Press, 1959.

______ **and Kessel, Reuben**, "Competition, Monopoly, and the Pursuit of Money," in H. G. Lewis et al., *Aspects of Labor Economics*, Princeton: Princeton University Press, 1962.

Ashenfelter, Orley, "Racial Discrimination and Trade Unionism," *Journal of Political Economy*, May–June, 1972, *80*, 435–464.

Becker, Gary S., *The Economics of Discrimination*, Chicago: University of Chicago Press, 1957, 1971.

Freeman, Richard, "Black Economic Progress After 1964: Who Has Gained and Why?," mimeo., National Bureau of Economic Research, 1978.

Friedman, Milton, *A Theory of the Consumption Function*, Princeton: Princeton University Press, 1957.

Griliches, Zvi, "Wages and Earnings of Very Young Men," *Journal of Political Economy*, No. 2, August 1976, *84*, S69–S86.

Gronau, Reuben, "Leisure, Home Production, and Work—The Theory of the Allocation of Time Revisited," *Journal of Political Economy*, December 1977, *85*, 1099–1123.

Gwartney, James, "Discrimination and Income Differentials," *American Economic Review*, June 1970, *60*, 396–408.

Hanoch, Guy, "An Economic Analysis of Earnings and Schooling," *Journal of Human Resources*, Summer 1967, 310–29.

Hashimoto, Masanori and Kochin, Levis, "A Bias in the Statistical Estimation of the Effects of Discrimination," *Economic Inquiry*, July 1980, *18*, 478–86.

Jenks, Christopher et al., *Inequality: A Reassessment of the Effect of Family and Schooling for America*, New York: Harper and Row, 1972.

Landes, William, "The Economics of Fair Employment Laws," *Journal of Political Economy*, July-August 1968, *76*, 507–53.

Pollak, Robert A., "Conditional Demand

Functions and Consumption Theory," *Quarterly Journal of Economics*, February 1969, *83*, 60–78.

Silberberg, Eugene, *Structure of Economics*, New York: McGraw-Hill, 1978.

Stigler, George, "The Economics of Information," *Journal of Political Economy*, June 1961, *69*, 213–25.

Weiss, Leonard and Williamson, Jeffrey G., "Black Education, Earnings and Interregional Migration: Some New Evidence," *American Economic Review*, June 1972, *62*, 372–83.

Welch, Finis, "Labor Market Discrimination: An Interpretation of Income Differences in the Rural South," *Journal of Political Economy*, June 1967, *75*, 225–40.

______, "Black-White Returns to Schooling," *American Economic Review*, December 1973, *63*, 893–907.

Testing Noncooperative Bargaining Theory: A Preliminary Study

By K. BINMORE, A. SHAKED, AND J. SUTTON*

Bargaining theory has received much attention of late. There has also been a growing interest in experimental work on bargaining, notably by Reinhard Selten (1978), and by Alvin Roth, M. Malouf, and J. Murnighan (1981). This work confirms a view that is common among social psychologists: namely, that subjects tend to seek a "fair" outcome to bargaining problems. The thrust of the inquiry is then to determine what the subjects will regard as fair in a given situation.

A tension exists between this work and the theoretical approach revitalized by Ariel Rubinstein (1982). (See also Binmore, 1982, 1983; Shaked and Sutton, 1984.) This new approach involves modeling the process of offer and counteroffer by means of which agreement can be reached, as a formal noncooperative game; and studying agreements that can be sustained as equilibria of this game.

The tension is sharply illustrated by a recent experimental study of W. Güth, R. Schmittberger, and B. Schwarze (1982). (See also Güth, 1983.) Two subjects have to divide a sum of money (the "cake"), using the following primitive procedure: Player 1 makes a demand, which Player 2 can then accept or refuse. This concludes the game. If the demand is refused, both players receive nothing. A strategic analysis assigns all (or nearly all) of the cake to Player 1, but experiments show that a much "fairer" division is usual.

The work of Güth et al. seems to preclude a predictive role for game theory insofar as bargaining behavior is concerned. Our purpose in this note is to report briefly on an experiment that shows that this conclusion is unwarranted. (Only the briefest account of the experiment is offered here; for a full account, see our 1984 paper.)

This does not mean that our results are inconsistent with those of Güth et al. Under similar conditions, we obtain similar results.[1] Moreover, our full results would seem to refute the more obvious rationalizations of the behavior observed by Güth et al. as "optimising with complex motivations." Instead, our results indicate that this behavior is not stable in the sense that it can be easily displaced by simple optimizing behavior, once small changes are made in the playing conditions.

I. The Experiment

In the present work, we went beyond the one-stage "ultimatum" game of Güth et al. and examined a two-stage game, as follows:

Stage I: The cake is of size 100 pence. Player 1 makes a proposal (X); Player 2 accepts (1 receives X, 2 receives $100-X$) or rejects (game continues).

Stage II: The cake is of size 25 pence. Player 2 makes a proposal (X'); Player 1 accepts (1 receives X', 2 receives $25-X'$) or rejects (1 receives 0, 2 receives 0).

A game-theoretic analysis requires that Player 1 makes an opening demand in the range 74–76 pence, and Player 2 accepts any opening demand of 74 pence or less (for he cannot do better by refusing, even if he obtains the entire cake in the second stage).

*London School of Economics, Houghton St., London WC2A 2AE, UK. We gratefully acknowledge the financial support of the International Centre for Economics and Related Disciplines (Suntory-Toyota Foundation), and the hospitality of the Psychology Department at LSE, where our experimental work was conducted, under the immediate supervision of Yasmin Batliwala, Mimi Bell, and Maria Herrero. We also thank Werner Güth, Alvin Roth, and particularly Reinhard Selten for comments on an early draft.

[1] See fn. 2 below.

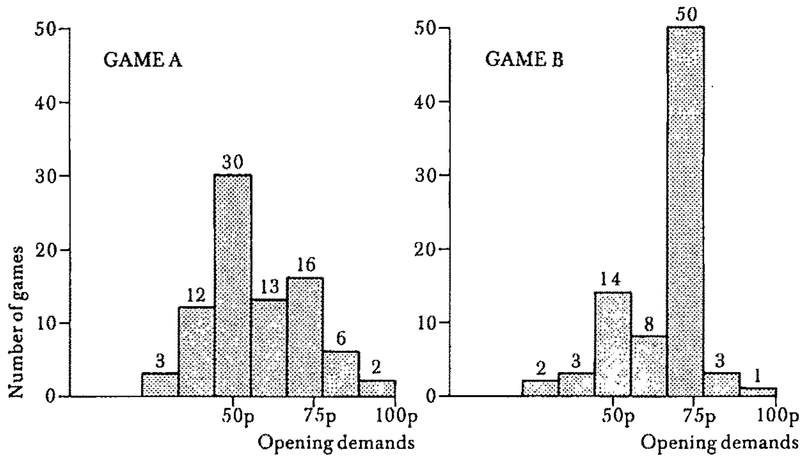


FIGURE 1. OPENING DEMANDS FOR MAIN RESULTS

We studied the game, using subjects who were isolated from each other, and who communicated their decisions via linked microcomputers. Following lengthy pilot studies, in which we solicited players' comments after they had played the game, we decided to extend the design, as follows. We invited the subject who had filled the role of Player 2 to play the game again, but this time he would fill the role of Player 1. We recorded only his opening demand in this second game (Game B).

II. The Results

We focus here on the main features of interest. The opening demands made in Game A and Game B, respectively, are shown in the histograms in Figure 1. They exhibit a marked change of behavior between Game A and Game B. A tendency to "play fair" in Game A becomes a strong tendency to play "like a game theorist" in Game B.

This marked change in behavior constitutes the first of the two main findings of the present study. The null hypothesis is that the opening demands in each game are drawn from the same population and is rejected at the 0.1 percent level (Kolmogoroff-Smirnoff two-tailed test).

Focusing on those subjects who filled the role of Player 2 in Game A, we looked at the subsample who faced a "high" demand in that game. A fair player would reject a high demand, and would not himself make a high demand (when offered the chance to act as Player 1, in Game B). *The results* (shown in Table 1) *indicate little support for the view that a substantial proportion of the population are "fairmen" as opposed to "gamesmen."* The table shows the relationship between a subject's response to the opening demand made of him in Game A, and the opening demand which he later makes when acting as Player 1 in Game B. Cell G denotes Gamesmen, cell F denotes Fairmen. We chose the midpoint between 50 and 75 as our dividing line between low and high demands. The table refers only to the subsample of our population who faced high demands in Game A.

TABLE 1

Opening Demand in Game B	Response to High Opening Demands ($63 \leq a \leq 77$) in Game A: No	Yes
$b \leq 62$	1(F)	2
$b \geq 63$	2	17(G)

What, then, of the players who filled the role of Player 1 in Game A, and who exhibited a marked tendency to make fair de-

mands? While we have considered various possible explanations, the interpretation that we favor is this: subjects, faced with a new problem, simply choose "equal division" as an "obvious" and "acceptable" compromise —an idea familiar from the seminal work of Thomas Schelling (1960). We suspect, on the basis of the present experiments, that such considerations are easily displaced by calculations of strategic advantage, once players fully appreciate the structure of the game.

Finally, it is important to note that Güth et al. did in fact study subjects playing the one-stage ultimatum game for a second time, without observing any marked change in behavior.[2] Thus, it is not *only* this feature which distinguishes our results from theirs.

The key feature to note, in this respect, is that *responses to opening demands in Game A* were strongly biased in favor of "rationality." (Of 22 opening demands in the range $63 \leq a < 77$, only 3 were rejected.) On the other hand, at the second stage of Game A —following a refusal at the first stage—subjects showed a strong tendency to reject high demands (as in the study of Güth et al).

Our suspicion is that the one-stage ultimatum game is a rather special case, from which it is dangerous to draw general conclusions. In the ultimatum game, the first player might be dissuaded from making an opening demand at, or close to, the "optimum" level, because his opponent would then incur a negligible cost in making an "irrational" rejection. In the two-stage game, these considerations are postponed to the second stage, and so their impact is attenuated.[3]

[2] Opening demands were slightly higher, and refusals of these demands *more frequent*.

[3] There remains the possibility that the difference between our results and those of Güth et al. might be traced to differences in the experimental environment rather than to differences in the game played. Güth et al. operated in an open environment within which subjects could see each other (although the identity of their current opponent was, of course, a secret). Our assistant, Yasmin Batliwala, has run a controlled experiment to check for this possibility (which will be reported separately). Replicating our experimental conditions, she compared the behavior of subjects playing our two-stage game with that of a control group playing the one-stage ultimatum game. Broadly, the results confirmed our present interpretation. Behavior in the two-stage game was similar to that reported in this paper. Behavior in the one-stage ultimatum game was consistent with the observations of Güth et al. in that game theory was a poor predictor of outcomes.

REFERENCES

Binmore, K. G., "Perfect Equilibria in Bargaining Models," ICERD Discussion Paper, London School of Economics, 1982.

______, "Bargaining and Coalitions I," ICERD Discussion Paper, London School of Economics, 1983.

______, **Shaked, A. and Sutton, J.,** "Fairness or Gamesmanship in Bargaining: An Experimental Study," ICERD Discussion Paper, London School of Economics, 1984.

Güth, W., "Payoff Distributions in Games and the Behavioral Theory of Distributive Justice," mimeo., Koln, 1983.

______, **Schmittberger, R. and Schwarze, B.,** "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization*, 1982, *3*, 367–88.

Roth, Alvin, Malouf, M. and Murnighan, J., "Sociological versus Strategic Factors in Bargaining," *Journal of Economic Behavior and Organization*, 1981, *2*, 153–77.

Rubinstein, Ariel, "Perfect Equilibrium in a Bargaining Model," *Econometrica*, January 1982, *50*, 97–109.

Shaked, A. and Sutton, J., "Involuntary Unemployment as a Perfect Equilibrium in a Bargaining Model," *Econometrica*, November 1984, *52*, 1351–64.

Selten, Reinhard, "The Equity Principle in Economic Behavior," in H. Gottinger and W. Leinfellner, eds., *Decision Theory and Social Ethics, Issues in Social Choice*, Dordrecht: Reifel Publishing, 1978, 289–301.

Did the United States Transmit the Great Depression to the Rest of the World?

By GERTRUD M. FREMLING*

This paper challenges the commonly held belief that the Great Depression was transmitted from the United States to the rest of the world. The well-known argument by Milton Friedman and Anna Schwartz (1963) is used as the reference point. I argue that their description, although intuitively plausible, does not correspond well with the data.

During the depression years and afterwards there was an extensive analysis of the international financial situation. A comprehensive analysis was provided by Ragnar Nurkse (1944). Earlier writings do not in particular indict the United States, but rather also blame France and several other countries for the financial crisis.[1] Recent literature that deemphasize the role of the United States include Peter Temin (1976) and Knut Borchardt (1982). Friedman and Schwartz do not get into any lengthy international analysis, yet draw far-reaching conclusions about the United States and the rest of the world.

This paper does not attempt to estimate the relative roles of French gold accumulations, the illiquidity of loans to Germany, the Austrian banking crisis, the British devaluation, etc. Rather the analysis is confined to evaluating whether the United States transmitted the depression to the rest of the world through the channels discussed by Friedman and Schwartz.

I. The Friedman-Schwartz Argument, and How It Can Be Evaluated

In their study (pp. 360–61), Friedman and Schwartz refer to several facts as evidence that, during the period August 1929 to August 1931, the United States transmitted the depression to the rest of the world. They refer to U.S. gold stocks, the flow of gold, whether the United States adhered to gold standard rules, and to the balance of payments.

The inflow of gold to the United States, the increase of U.S. gold stocks (= reserves?), and the deviation from gold standard rules are taken to imply that other countries were being forced to adapt to U.S. monetary policies. Apparently, Friedman and Schwartz believe that the United States exported the depression by lowering reserves in the rest of the world, leading to contractions of the money stock elsewhere. To evaluate this reserve argument, it is however necessary to study the data for the rest of the world, not —as Friedman and Schwartz do—merely for the United States. The test should be whether the rest of the world as a whole experienced a fall in gold reserves (alternatively, all international reserves including other foreign reserves). Gold flows to the United States, or increases in U.S. gold reserves, do not necessarily imply falling gold reserves elsewhere, since mining of gold as well as conversion of existing private gold stocks into currency can increase total world reserves.[2] Section II will

*Assistant Professor, University of Houston, University Park, TX 77004. I am grateful to Michael Darby, Paul Evans, and an anonymous referee for useful comments and suggestions. I also thank Lutz Spannagel, my research assistant, for double checking all calculations.

[1] For a sharp and thought-provoking criticism of French financial policies, see Paul Einzig (1932a, b; 1935; 1937).

[2] Gold flows generally do not equal changes in gold reserves, because gold flows measure movements of gold between countries not only of reserves, but also of private gold stocks. Further, changes in reserves include changes due to domestic redemption/minting, as well as international reserve flows. An outflow of gold from the rest of the world to the United States is therefore not

evaluate the correctness of the reserve argument. It will also discuss the related question of whether the United States followed gold standard rules to a smaller or greater extent than the rest of the world.[3]

There are two ways the U.S. balance of payments can affect the rest of the world: through reserves, and through real variables. For a surplus country, reserves are flowing in, draining reserves from other countries. The reserve question is, however, treated separately here. It is not clear whether Friedman and Schwartz are concerned solely with the influence through reserves, or also regard the direct real effect as important. Since the real effect, however, is a generally recognized mechanism, I also analyze this channel. By definition, the balance of trade, exports minus imports (which due mainly to the capital account does not equal the balance of payments), always equals the negative of the balance of trade for the rest of the world. Since it is a component of *GNP*, an increase in the U.S. trade surplus (or alternatively a decrease in the trade deficit) would have a restrictive influence. Such a scenario could occur if the depression originated in the United States. If income in the United States fell, imports would go down, and the lowered prices would stimulate exports. Notice that the issue is not whether the United States has a trade surplus, or trade deficit, or whether that surplus or deficit is large in absolute terms. A constant surplus, for instance, would fail to generate a swing in aggregate demand. Instead, the surplus or deficit must be related to surpluses or deficits in previous years. If the United States did transmit the depression to the rest of the world through the balance of trade, we should see a substantial "improvement" in the balance of trade.[4] The evidence is evaluated in Section III.

II. Evidence on Gold Reserves

According to data from the *Federal Reserve Bulletin*, June 1933, gold reserves in the rest of the world increased during the first two years of the depression, which is contradictory to the Friedman-Schwartz argument. For the period chosen by Friedman and Schwartz, the end of August 1929 to the end of August 1931, U.S. reserves increased from \$3.995 billion to \$4.632 billion; and total reserves for the world (fifty countries) increased from \$10.250 billion to \$11.297 billion. Hence, gold reserves outside the United States increased from \$6.255 billion to \$6.665 billion.

Further, this result is not sensitive to slight alterations in the time period considered. Calculations for each month between August 1929 to August 1931 reveal that there were only a few short periods during which gold reserves outside of the United States declined: March and April 1930, and January 1931. These declines were rather minor. The period August 1929 to August 1931 is actually more favorable to the Friedman-Schwartz argument than many other possible time periods. Between these dates, U.S. gold reserves were at least growing at a faster rate than gold reserves in the rest of the world, which was not the general trend. Using Federal Reserve data, Charles Hardy (1936, p. 93) calculated the shares of world gold reserves in the United States and in other countries. His table reveals that between December 1929 and December 1931, the U.S. share declined from 37.8 to 35.9 percent. A longer time span, December 1927 to Decem-

inconsistent with increasing reserves abroad. According to the League of Nations *Statistical Yearbook* data, the gold inflow (available yearly only) to the United States was \$120 million in 1929 and \$278 million in 1930. Friedman and Schwartz avoid applying their gold flow argument to the years 1927, 1928, and 1931, during which there were *outflows* of gold (1927, \$154 million; 1928, \$272 million; 1932, \$176 million). Had they done so, they could have concluded that Germany (whose depression started about a year earlier than in the United States) exported the depression to the United States, because it had a large gold inflow in 1928.

[3] Whether a country follows the rules is not by itself a criterion for evaluating whether a depression was exported or imported.

[4] Equal falls in exports and imports could possibly also have macroeconomic effects. Even though aggregate demand would not be directly affected, relative changes in the demand for different goods could conceivably affect total output in other countries.

ber 1933, shows an even greater decline: from 41.6 to 33.6 percent.

One might argue that since many central banks held foreign reserves (mostly balances with the Federal Reserve or the Bank of England), the measure should include foreign reserves in addition to gold. This is not obvious, however, since what is a foreign reserve asset for one country is an equally sized liability for another country. Under a gold standard the balances could be redeemed for gold, and the reserve countries would therefore need larger gold reserves in case of withdrawals. It turns out that for the period August 1929 to August 1931, the treatment of foreign reserves makes little difference. For twenty-eight countries (outside of the United States and the United Kingdom, which were the reserve centers),[5] the value of foreign reserves decreased from \$2.782 billion to \$2.559 billion, or by \$.223 billion. Since this is less than the \$.410 billion increase of gold reserves outside the United States, I still reach the same conclusion, that reserves in the rest of the world were increasing.

Even though it was shown above that the reserves in the rest of the world did not fall, the United States could be criticized for its policies if it was found that the Federal Reserve did not follow the "gold standard rules." Friedman and Schwartz claim that the United States not only violated these rules, but even went beyond 100 percent sterilization of gold flows:[6] "our money stock moved perversely, going down as the gold stock went up" (p. 361). While their statement regarding gold reserves and $M1$ is correct, they make the mistake of applying today's concept to earlier periods. The gold standard rules had no bearing on deposits in commercial banks, the main component of $M1$. The $M1$ measure was not a standard concept in those days, and "currency" and "credit" were usually referred to separately. For most nations, monthly data on currency were available, but information on credit was sketchy. In an evaluation, currency or high-powered money ought to be used.[7,8] It is true that while gold reserves increased by 15.9 percent, the currency stock only increased by 4.0 percent.[9] This does represent substantial sterilization since the absolute increase in currency (.193 billion) was less than the absolute increase in reserves (.637 billion).[10] Yet, these figures do not appear as strikingly bad when compared to the rest of the world. As an aggregate, it went beyond 100 percent sterilization: gold reserves increased by 6.6 percent, and currency stocks decreased by 4.5 percent.[11,12,13] Hence, I conclude that the

[5] Data on foreign assets were taken from the *Monthly Bulletin of Statistics*, various issues. Exchange rate data came from this source as well as from *Banking and Monetary Statistics 1914–1941* by the Federal Reserve. The 28 countries are: Albania, Austria, Belgium, Bulgaria, Canada, Chile, Czechoslovakia, Denmark, Egypt, Estonia, Finland, France, Germany, Greece, Hungary, India, Italy, Latvia, Lithuania, the Netherlands, Norway, Peru, Poland, Rumania, South Africa, Spain, Sweden, and Switzerland.

[6] If a central bank sterilizes reserve flows, domestic credit is decreased when reserves go up, and the inflow of reserves is prevented from increasing high-powered money by the full amount of the inflow. Since the money stock does not increase as much as otherwise (and prices do not adjust upwards as rapidly), the inflow of reserves may continue for an extended period. Sterilization by surplus countries can therefore cause prolonged drains of reserves from deficit countries.

[7] To follow the gold standard rules (often referred to as "the rules of the game"), it was required that a central bank increase its domestic assets when there was an inflow of gold, and decrease its domestic assets when there was an outflow. This would thus be the "opposite" of sterilization. A "strong" version of the rules required that the change in domestic assets be large enough to keep constant the "gold-reserve ratio" (alternatively defined as gold to currency ratio, gold to credit base ratio, or gold to liability ratio). For a discussion of the rules, see Nurkse (pp. 66–67) and Kenneth Dam (1982, pp. 17–18).

[8] Since figures for high-powered money are not available for most nations besides the United States, I report the currency stock figures.

[9] Using the figures in Friedman and Schwartz (Table A-1, col. (1), plus Table A-2, col. (1)). High-powered money increased by 3.1 percent (from Table B-3, col. (1)).

[10] High-powered money increased by .220 billion.

[11] Nurkse discusses the problem of worldwide sterilization during the interwar period. In his table on p. 69, which comprises 26 countries, he shows that sterilization was actually the most common behavior.

[12] The gold figures are the ones mentioned earlier in the text, based on the *Federal Reserve Bulletin*, June

United States did not violate the gold standard rules—as defined in those days—to any greater extent than most nations. The U.S. money stock—as defined by $M1$—moving perversely from August 1929 to August 1931 should not be blamed on abnormal sterilization policies, but rather on changes in the money multiplier.[14]

III. Evidence on the Balance of Trade

Did the U.S. balance of trade "improve" during 1929–31? The figures for exports minus imports are:[15] 1929: +\$842 million, 1930: +\$782 million, 1931: +\$334 million. Since the U.S. balance of trade showed a surplus every year in the preceding ten-year period, with an average of +\$1,380 million per year,[16] the United States cannot possibly have generated a downswing in the rest of the world through the balance of trade.

It could be argued that the quantities of exports and imports are more relevant than their values. Using quantities does however not change the results much. From 1928 to 1931, the export quantity index went down by 31 percent, and the import quantity index went down by 15 percent.[17] Taking a longer period, 1919 to 1931, gives a decline in the quantity of exports by 26 percent, and an increase in the quantity of imports by 20 percent.

IV. Conclusions

Analysis of data on countries other than the United States contradicts the hypothesis that the Great Depression was transmitted from the United States to the rest of the world during August 1929–August 1931 through the mechanism mentioned by Friedman and Schwartz. The figures also show that the United States did not deviate more from the gold standard rules than did most other countries during this period. I conclude that to the extent the Great Depression was transmitted internationally, other countries as well as the United States must have played a significant role, and/or it was transmitted from the United States through some channel not analyzed here.

Although this paper disagrees with the Friedman and Schwartz view of the international transmission, the evidence does not contradict other aspects of their description of the Great Depression. Neither does it rule out the possibility that the United States could have prevented or mitigated the worldwide depression through appropriate monetary policies.[18]

1933, subtracting U.S. figures from the world total. Data on currency stocks in 33 countries were taken from the League of Nations *Monthly Bulletin of Statistics*. To derive a rest-of-the-world aggregate currency stock, I multiplied each nation's currency stock with the appropriate exchange rate. Data on exchange rates for the two dates were found in the last mentioned source combined with *Banking and Monetary Statistics 1914–1941*. The 33 countries are: Australia, Argentina, Austria, Belgium, Brazil, Bulgaria, Canada, Chile, Czechoslovakia, Denmark, Egypt, Finland, France, Germany, Greece, Hungary, India, Italy, Japan, the Netherlands, Norway, New Zealand, Peru, Poland, Portugal, Rumania, South Africa, Spain, Sweden, Switzerland, the U.K., Uruguay, and Yugoslavia.

[13] Gold plus foreign reserves increased by less, 2.1 percent.

[14] According to the series presented in Friedman and Schwartz (Appendix B, Table B-3), the ratio of deposits to bank reserves fell from 13.09 to 11.69, and the ratio of deposits to currency held by the public from 10.81 to 8.95.

[15] According to the Census Bureau, *Historical Statistics of the United States*, Series U14.

[16] The figures immediately preceding the depression are slightly lower: 1928: +1,037 million, 1927: +681 million.

[17] Census Bureau, Series U225 and U237.

[18] There is no doubt that the United States possessed very large gold reserves in absolute terms, which suggests that an expansionary policy could have been pursued without invoking speculation of a devaluation.

REFERENCES

Borchardt, Knut, *Wachstum, Krisen, Handlungsspielräume der Wirtschaftspolitik*, Göttingen: Vandenhoeck & Ruprecht, 1982.

Dam, Kenneth W., *The Rules of the Game*, Chicago: University of Chicago Press, 1982.

Einzig, Paul, (1932a) *Finance and Politics*,

London: MacMillan, 1932.

______, (1932b) *Behind the Scenes of International Finance*, London: MacMillan, 1932.

______, *Bankers, Statesmen and Economists*, London: MacMillan, 1935.

______, *World Finance*, New York: MacMillan, 1937.

Fremling, Gertrud M., "Monetary Policy Under a Gold Standard: How Effective is It?," unpublished dissertation, University of California, December 1983.

Friedman, Milton and Schwartz, Anna J., *The Monetary History of the United States, 1867–1960*, Princeton: Princeton University Press, 1963.

Hardy, Charles O., *Is There Enough Gold?*, Washington: The Brookings Institution, 1936.

Nurkse, Ragnar, *International Currency Experience: Lessons of the Interwar Period*, Geneva: League of Nations, 1944.

Temin, Peter, *Did Monetary Forces Cause the Great Depression?*, New York: W. W. Norton, 1976.

Board of Governors of the Federal Reserve System, *Banking and Monetary Statistics 1914–1941*, Washington, D.C., 1943.

______, *Federal Reserve Bulletin*, Washington, various years.

U.S. Department of Commerce, Bureau of the Census, *Historical Statistics of the United States-Colonial Times to 1970*, Washington: USGPO, 1975.

League of Nations, *Monthly Bulletin of Statistics*, Publications Department, Geneva.

______, *Statistical Yearbook*, Economic Intelligence Service, Geneva.

Technology and Hedging Behavior: A Proof of Hicks' Conjecture

By MAUREEN O'HARA*

The idea that technology could lead to an imbalance in long and short hedging is due to John Hicks. He noted that:

> Technical conditions give the entrepreneur a much freer hand about the acquisition of inputs (which are largely needed to start a new process) than about the completion of outputs (whose process of production—in the ordinary business sense—may be already begun). If forward markets consisted entirely of hedgers, there would always be a tendency for a planned weakness on the demand side; a smaller proportion of planned purchases than of planned sales would be covered by forward contracts. [1946, p. 137]

It was this imbalance, Hicks conjectured, that would induce normal backwardation in forward prices.[1]

This paper provides a proof of Hicks' hedging conjecture. I demonstrate that, because of the properties of production functions and profit functions, long (or output) hedging may exceed short (or input) hedging *even if input and output decisions can be made at the same time*. The intuition behind this result is that the shape of a firm's technology may provide partial protection from price risks, but that this protection need not be symmetric with respect to input and output prices. To prove Hicks' conjecture, I develop a measure of the technology's price-risk sensitivity; this measure is similar to the Pratt-Arrow risk-aversion measure from utility theory. I then use this measure to indicate when producers will hedge input and output price-risks, and to provide sufficient conditions for Hicks' imbalance conjecture to hold.

That the technology per se, and not merely its timing, can lead to normal backwardation is the main contribution of this paper. My analysis can be viewed as a complement to the more extensive analyses of Hendrick Houthakker (1968), David Newbery and Joseph Stiglitz (1981), and Ronald Anderson and Jean-Pierre Danthine (1983) on the existence of normal backwardation. Those authors demonstrate that forward prices can be influenced by a wide variety of factors such as storage, timing, and quantity uncertainty. My analysis suggests that technological conditions can also be important in influencing financial sector behavior.

I. Production, Profit, and Prices

To focus on the effect of technology on hedging, I consider a very simple model of the economy. I assume that production is nonstochastic, so a firm (or producer) faces no quantity uncertainty in output.[2] However, each producer does face price uncertainty as the output price, p, and the input price, w, both depend on the state of the world $\theta, \theta \in [\underline{\theta}, \bar{\theta}]$. In the analysis that fol-

*Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853. I thank Lawrence Blume, David Easley, and Robert Frank for helpful comments. I also thank the National Science Foundation, grant no. IST-8408770, for financial support.

[1] Normal backwardation occurs when the forward price is less than the expected future spot price. Conversely, if the forward price exceeds the expected future spot price, this is referred to as contango.

[2] The case of stochastic production is extensively analyzed in Anderson and Danthine. They demonstrate that quantity uncertainty can result in either normal backwardation or contango in forward prices. This leads them to conclude "it is not possible from a purely theoretical point of view to demonstrate conclusively the predominance of backwardation or contango" (p. 388). In this paper I demonstrate that with *nonstochastic* production this predominance of backwardation can occur.

lows, I consider when a producer would prefer to "lock in" a price (either input or output) by entering a forward contract, or to wait and accept the uncertain price that will prevail in the spot market. Following Hicks, I define selling output via a forward contract to be long hedging; purchasing inputs via a forward contract to be short hedging.

Technology is given by $Y = g(X, Z)$ where Y is output, g is a production function, X is a variable input, and Z is a fixed input.[3] Initially, suppose that producers maximize expected profit. Then a producer selects X to maximize the expectation of the profit function

$$(1) \quad \pi(p,w) = p(\theta)g(X,Z) - w(\theta)X - Z$$

where the fixed input price is normalized to one. The following proposition characterizes the properties of this profit function.

PROPOSITION 1: *Let $g(X,Z)$ be strictly concave in X with $g'(X) > 0$, $g''(X) < 0$, and $g(0,Z) = 0$. Then*

(i) $\pi(p,w)$ is linear in p and linear in w if production decisions are made prior to the realization of $p(\theta), w(\theta)$.

(ii) $\pi(p,w)$ is strictly convex in p and strictly convex in w if production decisions are made subsequent to the realization of $p(\theta), w(\theta)$.

PROOF:

Given the timing of part (i), X is a number so $\pi(\cdot)$ is clearly linear in p and w. Part (ii) follows directly from Hal Varian (1978; Section 1.9).

Proposition 1 illustrates the relationship between prices and a producer's profitability. If producers cannot know the realizations of $p(\theta)$ and $w(\theta)$ before they produce, then (i) indicates that profit is linearly affected by price changes. This occurs because the producer's input and output amount is fixed; profit can vary only with price levels. If $p(\theta)$ and $w(\theta)$ are known, however, this simple relationship changes. Now the optimal amount to produce for each price realization can be selected; profit will vary because of quantity changes. As (ii) indicates, this quantity shifting means that different price levels have different, nonlinear effects on profit. The profit function becomes convex in input and output prices. The degree of convexity depends on the degree of concavity in the production function: the more concave the production function, the less convex the profit function.

The shape of the profit function has important implications for producers' forward market decisions. In case (i), the linearity of $\pi(p,w)$ means that an expected profit-maximizing producer would be indifferent between trading in the spot market at prevailing prices or trading in the forward market at the expected spot price. Since $\pi(p,w)$ is linear in both input and output prices, there is a symmetry in the forward demand to lock in these prices. This symmetry necessitates that the only forward market equilibrium has the forward price equal to the expected future spot price.

If the profit function is convex in prices, however, this is no longer the forward market equilibrium. The convexity of $\pi(p,w)$ means that producers actually *prefer* price variability. To induce producers to sell output forward, the forward price would have to exceed the expected future spot price. To induce producers to buy inputs forward, the forward price would have to be lower than the expected future spot price. No forward market composed only of hedgers could exist.

A forward market equilibrium characterized by normal backwardation, therefore, cannot occur if producers maximize expected profits. Since producers are, at worst, indifferent to price risk, they have no reason to "lock in" prices. This suggests that a necessary condition for Hicks' conjecture to hold is that (at least some) producers be averse to

[3] In the analysis that follows, the concavity of $g(\cdot)$ in X plays an important role. However, this needed concavity is with respect to a specific input, and not necessarily with respect to total inputs. For example, a Cobb-Douglas technology of the form $Y = X_1^a X_2^b$ with $a + b = 1$ is linear in total inputs but is concave in specific inputs. For simplicity, the paper concentrates on the single-variable input case, but the analysis can easily be adapted to the multiple input case.

price risks. The analysis demonstrates that this aversion will not arise from the technology; if it arises at all, it must be because of producer risk preferences.

II. Utility, Technology, and Hedging

If producers are risk averse in profit, then their objective function is given by $E[U(\pi(p,w))]$, where U is a strictly concave utility function. Although $U(\cdot)$ is concave in profit, whether it is also concave in prices depends upon the relative shapes of the utility and profit functions. The shape of the utility function can be characterized by its measure of absolute risk aversion. The following proposition suggests a similar measure to characterize the shape of the profit function. These two measures provide a general rule for when risk aversion with respect to profit also implies risk aversion with respect to prices. If producers are averse to price risks, they will prefer to lock in prices in the forward market rather than accept the variable prices in the spot market.[4]

PROPOSITION 2: *Let $\pi(p,w)$ be the profit function, and $U(\pi(p,w))$ be the objective function. Let U_π denote a first derivative, $U_{\pi\pi}$ a second derivative (similarly for π). Then*:

(*i*) *A producer will hedge input price risks if in the absence of hedging*

$$(2) \qquad -U_{\pi\pi}/U_\pi > (1/\pi_w)\pi_{ww}/\pi_w.$$

(*ii*) *A producer will hedge output price risks if in the absence of hedging*

$$(3) \qquad -U_{\pi\pi}/U_\pi > (1/\pi_p)\pi_{pp}/\pi_p.$$

[4]The decision to enter the forward market also depends upon the forward prices. In the analysis that follows, I evaluate forward demand and supply when the forward price is equal to the expected future spot price. If a producer's objective function is concave in prices, however, he would be willing to pay a premium to enter the forward contract. Hence, if long hedging exceeds short hedging, the forward price will fall, resulting in normal backwardation. If short hedging exceeds long hedging, the forward price will rise, resulting in contango.

PROOF:

Let $U(\pi(p,w)) \equiv V(p,w)$. Then to establish (2) note that $V_w = U_\pi \pi_w < 0$ and $V_{ww} = U_{\pi\pi}(\pi_w)^2 + U_\pi \pi_{ww} \gtrless 0$. Concavity, or risk aversion in w, requires $V_{ww} < 0$. This implies $U_\pi \pi_{ww} < -U_{\pi\pi}(\pi_w)^2$ and the result follows. To establish (3) note that $V_p = U_\pi \pi_p > 0$ and $V_{pp} = U_{\pi\pi}p(\pi_p)^2 + U_\pi \pi_{pp} \lesseqgtr 0$. Concavity, or risk aversion in p, requires $V_{pp} < 0$. This implies $U_\pi \pi_{pp} < -U_{\pi\pi}(\pi_p)^2$ and the result follows.

The general rule stated above gives a simple way to characterize producers' forward market decisions. If a firm is risk neutral in profit, then $U_{\pi\pi} = 0$ and, as expected, the firm will not hedge. If the firm is risk averse, it will hedge if its absolute aversion to risk overwhelms the inherent convexity of the profit function. In the case of a linear profit function, π_{ww} and π_{pp} are zero, so hedging becomes attractive. This will occur, for example, when the firm's production decisions are made in advance of price realizations. If the profit function is strictly convex, however, even a very risk-averse producer may prefer not to hedge price risks.

One implication of these conditions is that the timing of production may be sufficient to generate an imbalance in long and short hedging. If, as Hicks implies, input prices are more often known in advance of production than are output prices, then the profit function will be linear in p but convex in w. This means that producers will be more willing to hedge output prices than input prices. If only hedgers were present in forward markets, this imbalance would induce normal backwardation.

Timing differences, however, are not necessary to generate this hedging imbalance. The technology itself may be sufficient to induce differential hedging. This can be demonstrated by incorporating some additional properties of profit functions and risk aversion measures into the analysis. To avoid timing complications, I assume that both input and output prices are known before production.[5] Then, from Hotelling's Lemma,

[5]With both input and output prices known before production, the producer has the maximum amount of

$-\partial\pi(p,w)/\partial w = X$ and $\partial\pi(p,w)/\partial p = Y$ where X is the input demand and Y is output. Substituting for π_w and π_p in equations (2) and (3) and multiplying by π yields

$$(4) \qquad -U_{\pi\pi}\pi/U_\pi > (-\pi/X)\pi_{ww}/\pi_w,$$

$$(5) \qquad -U_{\pi\pi}\pi/U_\pi > (\pi/Y)\pi_{pp}/\pi_p.$$

The left side of the expressions is the producer's relative risk-aversion measure. From (4), the producer's desire to hedge input price risks depends upon the relationship of both profit per unit of input and the technology to this relative risk-aversion measure. Equation (5) draws a similar comparison with profit per unit of output. One implication of this latter condition is that the higher the profit per unit of output, the more likely is the firm to hedge output price risks.

Equations (4) and (5) illustrate the interrelatedness of the hedging decision with technology. This linkage can be understood better by examining how the technology and profit function relate. Write the technology as $Y = g(X(p,w), Z)$ where $X(\cdot)$ is the input demand function. The profit function is then $\pi(p,w) = pg(X(p,w), Z) - wX(p,w)$. It is easy to show that

$$(6) \qquad \pi_{ww} = pg''(X_w)^2 - 2X_w$$

$$(7) \qquad \pi_{pp} = pg''(X_p)^2 + 2g'X_p$$

where primes denote derivatives of $g(\cdot)$ and subscripts denote derivatives of $X(\cdot)$.

Equations (6) and (7) can be simplified by realizing that, from the first-order conditions, profit maximization requires $pg'(X(p,w)) - w = 0$. This implies that $pg''X_w = 1$ and $pg''X_p = -g'$. It follows that

$$(8) \qquad \pi_{ww}/\pi_w = 1/Xg''(\cdot)p$$

$$(9) \qquad \pi_{pp}/\pi = -(g'(\cdot))^2/g(\cdot)g''(\cdot)p.$$

It is now easy to see how the shape of the technology influences the hedging conditions given by Proposition 2. As the technology becomes more linear, $g''(\cdot)$ goes to zero from below and both profit function conditions go to infinity. With the right sides of equations (2) and (3) infinite, even an extremely risk-averse producer will not hedge. As before, the intuition behind this result is that the shape of the technology already provides insulation from price changes. The less concave this technology, the larger are the changes in X and Y for changes in p and w.

These price-induced changes in X and Y, however, need not be symmetric. As a result, the ability of the production process to limit input and output price risks may also not be symmetric. To see why, recall that $X_w = 1/pg''$ and $X_p = -g'/pg''$, and note that the output elasticity with respect to X, denoted η_X, is equal to $g'(\cdot)X/g(\cdot)$. Then equations (8) and (9) can be rewritten as

$$(10) \qquad \pi_{ww}/\pi_w = X_w/X$$

$$(11) \qquad \pi_{pp}/\pi_p = \eta_X(X_p/X).$$

The elasticity effect in equation (11) dictates that input and output prices affect the firm's production decisions in very different ways. While both price changes directly affect the input demand function, output price changes also have an indirect effect through the production function. The "warping" effect of the production function limits the technology's effectiveness in hedging output price risks. As the following proposition demonstrates, this implies that firms are more likely to hedge output price risks than input price risks.

PROPOSITION 3: *Given a strictly concave production function $g(X)$ with $g(0) = 0$,*

$$(12) \quad (1/\pi_p)\pi_{pp}/\pi_p < (1/\pi_w)\pi_{ww}/\pi_w.$$

PROOF:

From equations (8), (9), and Hotelling's Lemma, this condition can be rewritten as

$$(-(g')^2/pgg'')1/g < 1/Xpg''(-1/X).$$

production flexibility. This allows the producer to partially offset unfavorable price conditions, and thus reduces his incentives to hedge. By focusing on this case, the analysis examines the interaction between production flexibility and hedging when production flexibility is at its greatest. Certainly, if the producer must take production decisions before prices are known, his hedging decisions may differ.

Simplifying yields

$$(g'/g)(g'/g) < (1/X)(1/X)$$

which is equivalent to

$$(g'X/g)(g'X/g) < 1.$$

But $(g'X/g)$ is η_X. Hence, we need $(\eta_X)^2 < 1$ to establish the claim. Since $\eta_X > 0$, it is sufficient to establish $\eta_X < 1$. Thus we need $Xg'/g < 1$, or $g' < g/X$. But this is equivalent to Marginal Product of $X <$ Average Product of X, which is always true if $g(\cdot)$ is strictly concave and $g(0) = 0$.

Proposition 3 provides an important result. If a firm will hedge input price risks, it will also hedge output price risks. However, a firm may hedge output prices *without* wanting to hedge input prices. As a result, the demand for "output price insurance" exceeds the demand for "input price insurance." As Hicks' conjectured, the technology induces an imbalance in long and short hedging. If forward markets were composed only of hedgers, then normal backwardation would result.

III. Concluding Remarks

This paper has examined the effect of technology on producers' forward market decisions. I have demonstrated that the technology itself may provide protection from price risks. This technological protection obviates the need for forward market protection, particularly in the case of input prices. This results in the imbalance of long and short hedging that Hicks conjectured would characterize forward market participation.

While I have demonstrated sufficient conditions for Hicks' conjecture to hold, it is certainly true that in other economic environments this hedging imbalance may not arise. As Anderson and Danthine have demonstrated, the presence of quantity uncertainty or storage can introduce other factors into producers' hedging decisions. Nevertheless, by focusing strictly on technology, this paper demonstrates the crucial link between technology and hedging behavior.

REFERENCES

Anderson, Ronald W. and Danthine, Jean-Pierre, "Hedger Diversity in Futures Markets," *Economic Journal*, June 1983, *93*, 370–89.

Hicks, John R., *Value and Capital*, 2nd ed., London: Oxford University Press, 1946.

Houthakker, Hendrick S., "Normal Backwardation," in J. N. Wolfe, ed., *Value and Growth: Papers in Honor of Sir John Hicks*, Edinburgh: Edinburgh University Press, 1968.

Newbery, David M. G. and Stiglitz, Joseph E., *The Theory of Commodity Price Stabilization*, Oxford: Oxford University Press, 1981.

Varian, Hal R., *Microeconomic Analysis*, New York: W. W. Norton, 1978.

Equilibrium Price Dispersion

By WILLIAM HALLAGAN AND WAYNE JOERDING*

Steven Salop and Joseph Stiglitz (1982) present a model of equilibrium price dispersion which exhibits both single-price equilibria (*SPE*) and two-price equilibria (*TPE*).[1] A distinctive characteristic of this model is that, while *SPE* are possible at the high (monopoly) price, the Salop-Stiglitz model does not exhibit *SPE* at the low price. This result raises the question of whether *SPE* can exist at the competitive price in the presence of consumer search. We present a model which, in many ways, is similar to the Salop-Stiglitz model, but it is capable of exhibiting *SPE* at either the high or low price as well as *TPE*. Section I develops the model which will, to the extent convenient, follow the notation used in Salop and Stiglitz (1982).[2] After demonstrating that *SPE* at the high and low price are possible, we conclude with a discussion of the major difference between this model and Salop and Stiglitz.

I. The Model

Suppose there are a large number of firms (N) selling a homogeneous commodity and facing the same production technology. Firms are allowed to differ only in that some select the high price (P_h) strategy while others select the low price (P_l) strategy. For convenience, P_h and P_l are measured net of constant unit costs. Following Salop and Stiglitz, there exist a large number of consumers (L) who purchase one unit per period as long as the price is not greater than some reservation price u. Consumers purchase nothing from firms charging more than u. We assume that consumers are heterogeneous with differing propensities to search for low-price stores. Assume that consumer i will seek out the low-price store if the average price differential ($\overline{D}$) is greater than some constant d_i. If $f(\cdot)$ describes the distribution of d_i for $i=1,\dots,L$, then

$$(1)\qquad \alpha(\overline{D})=\int_0^{\overline{D}} f(x)\,dx$$

is the proportion of consumers seeking out the low-price firm. These consumers are referred to as type-*I* consumers, while we call the remaining $[1-\alpha(\overline{D})]L$ consumers type *II*.[3]

Type-*I* consumers search varying amounts and, as long as low-price firms exist, no type-*I* consumer will purchase anything from a high-price firm. Let $M_i(D_i,\overline{D})$ be firm i's share of the type-*I* consumers where it sets its price $P_i=P_h-D_i$ with $D_i>0$. If $D_i=0$ then $M_i(0,\overline{D})=0$. We assume that type-*II* consumers distribute themselves equally across all firms. The profit function for firms following the high-price strategy is

$$(2)\qquad \Pi_{hi}=\begin{cases} P_h L\left[\dfrac{1-\alpha(\overline{D})}{N}\right] & \text{for } P_h\le u\\ 0 & \text{for } P_h>u\end{cases}$$

Throughout the rest of the development, we assume that firms treat $\overline{D}$ and u as known

*Department of Economics, Washington State University, Pullman, WA 99164.

[1] Other papers treating equilibrium price dispersions are Salop and Stiglitz (1977), Jennifer Reinganum (1979), A. Braverman (1980), J. A. Carlson and R. P. McAfee (1983), and K. Burdett and K. L. Judd (1983).

[2] The model is a hybrid of the Salop and Stiglitz model and the models developed in our 1983, 1985 articles.

[3] Salop-Stiglitz emphasize that in their model consumers are identical with respect to incomes and preferences, but, by chance, some happen to arrive at high-price stores while others arrive at low-price stores. Braverman models consumers as identical except with respect to the costs of gathering market price information. Braverman concludes that the distribution of these search costs is critical in determining the nature of the market equilibrium.

parameters. An immediate consequence is

LEMMA 1: *The profit-maximizing price for firms following the high-price strategy is* $P_h = u$.

This is Lemma 3 of Salop-Stiglitz and follows from the maximization of (2) with respect to P_h. With $\overline{D}$ treated as a constant, profits are positive and increase with P_h until P_h exceeds u and profits drop to zero. Thus, profits are highest when $P_h = u$.

Lemma 1 is useful for both Salop-Stiglitz and this model because it allows us to use $P_h = u$ as a reference point for other prices. We express prices for firms following the low-price strategy as $P_{li} = P_h - D_i$. The profit function for firms following the low-price strategy is

$$(3)\quad \Pi_{li} = (P_h - D_i)L \times\left[\alpha(\overline{D})M_i(D_i,\overline{D}) + \frac{1-\alpha(\overline{D})}{N}\right].$$

Maximizing (3) with respect to D_i and rearranging the first-order condition the profit-maximizing D_i must satisfy (4).

$$(4)\quad D_i\left[\alpha(\overline{D}) + (1-\alpha(\overline{D}))/M(D_i,\overline{D})N\right] - (P_h - D_i)\alpha(\overline{D})E_i = 0$$

where $E_i = (dM/dD_i)(D_i/M)$ is the firm's expectation of the elasticity of its share of type-I consumers with respect to its choice of D_i. The E_i contains both the expected direct effect of its own price discount as well as the conjectured effect of its price discount on the price discount of other firms. E_i is assumed to be constant over the relevant range of price discounts.[4]

LEMMA 2: *If* $E_i = E$ *for all firms then an equilibrium will, at most, contain two prices;* $P_h = u$ *and* $P_l = u - D$, *where* $D = D_i$ *for all low-price firms and satisfies* (4).

Given the result of Lemma 1, this result follows from the derivation of (4). Hereafter, we assume the $E_i = E$ for all firms so that Lemma 2 justifies our reference to the low price without making allowances for multiple low prices. Furthermore, we will assume that with all low-price firms charging the same price, type-I consumers are shared equally by low-price firms so that $M_i(D_i, \overline{D}) = 1/N_l$ where N_l is the number of low-price firms.

This model can exhibit two types of equilibria, SPE and TPE. As discussed in the introduction, our concern is whether a model that exhibits a TPE can also exhibit both varieties of SPE; high-price SPE and low-price SPE. It is useful to define $\Delta(D)$ as the difference in profit between low-price firms and high-price firms from (2) and (3),

$$(5)\quad \Delta(D) = \Pi_l - \Pi_h = \frac{(P_h - D)L\alpha(D)}{N_l} - DL[1-\alpha(D)]/N.$$

PROPOSITION 1: *A necessary condition for the existence of a high-price SPE is* $E > 1$.

A high-price SPE will exist only if $\Delta(D^*) < 0$ when $N_l = 1$ and D^* is the profit-maximizing price discount where $N_l = 1$. Clearly $D^* \le P_h$ and $\alpha(D^*) \le \alpha(P_h)$, thus, for $N_l = 1$, $\Delta(D^*) < 0$ implies

$$(6)\quad (P_h - D^*)L\alpha(P_h) - D^*L[1-\alpha(P_h)]/N < 0.$$

But (6) can be rearranged to obtain

$$(7)\quad \frac{\alpha(P_h)N}{1-\alpha(P_h)} < \frac{D^*}{P_h - D^*}.$$

A lone low-price firm captures all type-I consumers so that $M(D^*, D^*) = 1$, and the maximum profit condition (4) can be expressed as

$$(8)\quad \frac{D^*}{P_h - D^*} = \frac{N\alpha(D^*)E}{N\alpha(D^*) + 1 - \alpha(D^*)}.$$

[4] Even though a constant E_i is likely to be inconsistent with the behavior of consumers and other firms, it is assumed that, as a rule of thumb, firms make their decisions as if E_i were a constant.

Noting that the right-hand side of (8) is increasing in α, we can substitute $\alpha(P_h)$ for $\alpha(D^*)$ and combine (7) and (8) to form

$$(9)\quad \frac{\alpha(P_h)N}{1-\alpha(P_h)} < \frac{N\alpha(P_h)E}{N\alpha(P_h)+1-\alpha(P_h)},$$

which can be rearranged as

$$(10)\quad E > \frac{1-\alpha(P_h)+N\alpha(P_h)}{1-\alpha(P_h)} > 1.$$

Thus, E must exceed 1 if an *SPE* at the high price exists.

Proposition 1 is counterintuitive. A value of E greater than 1 would imply that firms expect that total revenue will increase with a drop in price and one would think that firms would be more likely to adopt a price discount strategy in this case. Proposition 1, however, is about an equilibrium. It may be that such an *SPE* is likely to be unstable. It is interesting to note that while elasticities exceeding 1.0 often cause problems for equilibrium in models of promotional strategies (Richard Schmalensee, 1972), in this model $E>1$ may, in fact, be interpreted as a conjectural elasticity which is consistent with equilibrium. That is, if firms expect that their price cut will not be followed by other firms, then it is likely that they will consider E to exceed 1 and, for a high-price *SPE* with $E>1$, no firm will follow any other firm experimenting with a low-price strategy.

PROPOSITION 2: *If $N \geq 2$ a sufficient condition for a low-price SPE to exist is $E<1$.*

A low-price *SPE* will exist if $\Delta(D^*)>0$ when $N_l = N-1$ where D^* is the profit-maximizing price discount satisfying (4). From (5), $\Delta(D^*)>0$ when $N_l=N-1$ implies

$$(11)\quad (P_h - D^*)\alpha(D^*) > \frac{N-1}{N}D^*[1-\alpha(D^*)].$$

Rearranging (4), when $N_l = N-1$, and maintaining the assumption that $M(D^*, D^*) = 1/N_l$, D^* must satisfy

$$(12)\quad (P-D^*)\alpha(D^*) = \frac{D^*}{E}\left[\alpha(D^*)+\frac{N-1}{N}(1-\alpha(D^*))\right].$$

Combining (11) and (12) the condition for $\Delta(D^*)>0$ can be expressed as

$$(13)\quad \frac{N\alpha(D^*)}{(N-1)(1-\alpha(D^*))}+1>E.$$

If we are willing to adopt the simplifying assumption that α is a constant, Propositions 1 and 2 can be strengthened. A constant α may in fact be a reasonable assumption if we think that individual shopping behavior is formed on the basis of experiences with the price dispersions existing in the past, and that this behavior is either slow to adapt to current price dispersions or that price dispersions have not changed perceptibly. In the case of a constant α, Propositions 1 and 2 can be restated as Propositions 1A and 2A below.

PROPOSITION 1A: *If α is a constant, a necessary and sufficient condition for the existence of a high-price SPE is $E>1+N(\alpha/(1-\alpha))$.*

PROPOSITION 2A: *If α is a constant and $N\geq 2$, a necessary and sufficient condition for the existence of a low-price SPE is $E<1+(N/(N-1))(\alpha/(1-\alpha))$*

PROPOSITION 3: *If α is a constant, and $N\geq 2$ a necessary and sufficient condition for a two-price equilibrium to exist is*

$$1+\left(\frac{N}{N-1}\right)\left(\frac{\alpha}{1-\alpha}\right) \leq E \leq 1+N\left(\frac{\alpha}{1-\alpha}\right).$$

From Propositions 1A and 2A for values of E between $(1+(N/(N-1))(\alpha/(1-\alpha))$ and $(1+N(\alpha/(1-\alpha))$ there cannot be a single-price equilibrium at either the high price or the low price. To determine the range of E such that a *TPE* exists, we need to show that there exists a value of N_l such that $1 \leq N_l < N-1$ and $\Delta(D^*)=0$. The $\Delta(D^*)=0$ im-

plies

$$(14)\quad N_l = (P_h - D^*)\alpha N / D^*(1-\alpha).$$

Substituting into (14) for the profit-maximizing value of $((P - D^*)/D^*)$ from (4) for the case where α is a constant gives us

$$(15)\quad N_l = N\alpha/((E-1)(1-\alpha)).$$

Using (15), $1 \le N_l \le N-1$ when

$$\left(\frac{N}{N-1}\right)\left(\frac{\alpha}{1-\alpha}\right)+1 \le E \le N\left(\frac{\alpha}{1-\alpha}\right)+1$$

as claimed in Proposition 3.

II. Conclusion

The model presented by Salop and Stiglitz (1982) was incapable of exhibiting *SPE* where all firms charged the low price. This peculiar result is generated by the nature of demand facing their low-price firm. In the Salop-Stiglitz model, low-price firms faced perfectly inelastic demands for further price decreases, but if they raised price infinitesimally their sales would drop to the level of the high-price firms. In this setting, the profit-maximizing discount (using our terminology) for Salop and Stiglitz's low-price firm is

$$(16)\quad D^* = \delta(N/(N-N_l))$$

where δ is a positive constant.[5]

Given this background it is easy to see that a low-price *SPE* ($N_l = N$) will not exist in their model since as N_l approaches N, D^* approaches infinity, and the low-price firms charge infinitely negative prices, while firms that switch to the high-price strategy will experience positive profits. In the model presented in this paper, the demand facing the low-price firms is a smooth function of price, and, as a result, the model is capable of exhibiting *TPE* as well as *SPE* at the low and high price. In the simple case where α is a constant, the profit-maximizing discount for the low-price firm is

$$(17)\quad D^* = P_h\left[\frac{\alpha E}{\alpha + (1-\alpha)(N_l/N) + \alpha E}\right]$$

which, for positive E, is always less than P_h so that the profit-maximizing P_l is always positive.

[5] This is the result of the principal case treated by Salop and Stiglitz where the costs of entering the market (c) are zero and storage costs (δ) are positive. In this case they show, from their equation (9), that $P_l = u - \delta(N/(N-N_l))$.

REFERENCES

Braverman, A., "Consumer Search and Alternative Equilibria," *Review of Economic Studies*, April 1980, *47*, 487–502.

Burdett, K. and Judd, K. L., "Equilibrium Price Dispersion," *Econometrica*, July 1983, *51*, 955–70.

Carlson, J. A. and McAfee, R. P., "Discrete Equilibrium Price Dispersion," *Journal of Political Economy*, June 1983, *91*, 480–93.

Hallagan, W. and Joerding, W., "Polymorphic Equilibrium in Advertising," *Bell Journal of Economics*, Spring 1983, *14*, 191–201.

______ **and** ______, "Polymorphism in Non-Price Competition," *Journal of Business and Economics*, forthcoming 1985.

Reinganum, Jennifer F., "A Simple Model of Equilibrium Price Dispersion," *Journal of Political Economy*, August 1979, *87*, 851–58.

Salop, S. and Stiglitz, J. E., "Bargains and Ripoffs: A Model of Monopolistically Competitive Price Dispersion," *Review of Economic Studies*, October 1977, *44*, 493–510.

______ **and** ______, "The Theory of Sales: A Simple Model of Equilibrium Price Dispersion with Identical Agents," *American Economic Review*, December 1982, *72*, 1121–30.

Schmalensee, Richard, *The Economics of Advertising*, Amsterdam: North-Holland, 1972.

The Quality of Schooling: Comment

By PETER J. EATON*

A recent article in this *Review* by Jere Behrman and Nancy Birdsall (1983) raises an important issue for makers of educational policy in developing countries. Behrman-Birdsall break ground both in specifying a model that treats schooling quantity and quality simultaneously and systematically, and in developing practical methods for evaluating social rates of return to schooling quantity and quality in the absence of externalities. Their model is estimated using data from Brazil, a country with an impressive recent record in expanding primary schooling.

This comment has two objectives. The first is to point out methodological errors made by Behrman and Birdsall in the derivation of the social rate of return to schooling quantity. The second is to test the robustness of this (corrected) finding with a data set that differs from that used by Behrman and Birdsall.

The principal finding of the Behrman-Birdsall analysis is that the estimated social rate of return to schooling quantity in Brazil is much lower than that to schooling quality (6.8 vs. 10.4 percent). If this is generally the case in developing countries, the possibility of a conflict between productivity and equity goals becomes an issue to be taken into account by educational policymakers. For example, with the conclusions reached by Behrman-Birdsall, "If only one-third of the students were to attend school and the resources saved by schooling only one-third of the students could be used to increase both quality and quantity by 50 percent, the estimates imply an [overall] income gain of 32 percent!" (p. 941).

The analysis that follows demonstrates that correct procedure yields an estimated social rate of return to schooling quantity which is only slightly lower than that to schooling quality (9.8 vs. 10.4 percent), indicating no substantial inefficiency in the allocation of educational resources between schooling quantity and quality. To test the robustness of this result, a model of the Behrman-Birdsall type is estimated using a 1979 set of micro data from one city in the Brazilian northeast. Using theoretically and empirically preferred variable definitions, very similar estimation results are obtained with respect to relative social rates of return to quantity and quality.

Section I contains a correction of the errors made in deriving the social rate of return to schooling quantity. Section II is empirical and contains a summary of potential sources of bias, a description of the data set used to estimate a Behrman-Birdsall-type model, and estimation results. Section III contains a discussion of some remaining problems.

I. Correction of the Social Rate of Return to Years of Schooling

The preferred model developed by Behrman and Birdsall is a modification of the model developed by Jacob Mincer (1974). The difference between their model and the basic Mincerian specification lies in their inclusion of schooling quality. From their equations (2) and (2′), the model estimated can be summarized as

$$(1)\quad \ln Y_{S,Q} = \ln Y_0 + \left(r_0 + r_1 Q + r_2 Q^2\right) S + aE + bE^2 + W,$$

*Associate Professor of Economics, University of Missouri, 5100 Rockhill Road, Kansas City, MO 64110. The Brazilian Planning Ministry provided a grant which generated the empirical basis for this comment. I am grateful to David A. Denslow, Jr. for constructive insights with respect to substantial issues addressed here, and to the personnel at the Curso de Mestrado em Economia of the Federal University of Ceara, Brazil, for their guidance and help in gathering data. The views expressed are my own and do not necessarily reflect the opinions of any of these institutions or individuals.

where Y is monthly earnings, Q is a measure of schooling quality, S is years of schooling, E is labor force experience, and W is a random disturbance. Regression coefficients provide the information necessary to calculate an estimated social rate of return to years of schooling for a given quality, either as an average or marginal rate. Behrman and Birdsall use a method that yields an average rate. The marginal procedure is preferable, both for consistency with their computation of the social rate of return to schooling quality and for econometric considerations.[1]

The following procedure assumes that an individual with sample mean values of S, Q, and E behaves according to standard human capital theory with respect to investment in schooling quantity and quality, that the only private gain involved is that of an increased lifetime income, and that the only private cost involved is that of postponing entry to the labor force for an additional year.

If Y^* is the gross private gain in lifetime earnings resultant from an additional year of schooling, then the present value (G^*) of that gain can be written as

$$(2)\quad G^* = Y^* \int_1^{N+1} e^{-rt}\,dt = Y^*(-1/r)\left[e^{-r(N+1)} - e^{-r}\right].$$

Using the estimated coefficients from Behrman and Birdsall's *OLS* estimation of (1) yields $\ln Y$ of 4.63 for the average individual.[2] The gain in $\ln Y$ for this individual from an additional year of schooling is $\delta \ln Y/\delta S = .117$. Since earnings are measured in monthly terms, estimated Y^* is $12(e^{4.63+.117} - e^{4.63}) = 154$.[3] Assuming $N = 40$, equation (2) then yields $G^* = 1153$.

The present value of the private cost of an additional year of schooling ($C*$), which is equal to G^* at equilibrium, is[4]

$$(3)\quad C* = P^* \int_0^1 e^{-rt}\,dt = P^*(-1/r)(e^{-r} - 1),$$

where P^* is the annual cost of an additional year of schooling for the individual. Substituting for r and using the equilibrium condition ($G^* = C*$) results in $P^* = 1222$. Assuming that schooling is a constant cost industry over the relevant range, the Behrman-Birdsall estimate of the average annual social cost of schooling (242) can be used to approximate the social cost of an additional year of schooling, which is added to P^* to yield the total cost of an additional year of schooling. The social rate of return to years

[1] Ordinary least square (*OLS*) estimation of the semi-log form yields estimates of the change in Y resulting from a change in S (or Q) that are biased. In the analysis that follows, gains associated with schooling are expressed in terms of Y. The bias involved in estimating this change from a change in $\ln Y$ is smaller if the change in $\ln Y$ is associated with a marginal change in S than if it is associated with all schooling.

[2] The sample mean value of $\ln Y$ is 4.51 (p. 933). This value does not correspond to $\ln Y$ at mean values of S, Q, and E because of the squared terms in the quality and experience variables. Behrman and Birdsall calculate rates of return to both quantity and quality of schooling using 4.51 to estimate $\ln Y$ for the average individual.

[3] The principal reason for the Behrman-Birdsall finding of a low estimated rate of return to schooling quantity relative to quality is that they neglect to transform the monthly earnings to annual earnings in their calculation of the social rate of return to schooling quantity (p. 940, fn. 29). In calculating the total cost of an additional year of schooling they add average private cost, measured as average *monthly* foregone earnings, to average social cost, measured as a function of average *annual* teachers' salaries. Ignoring other problems, the correct procedure, rather than adding 304 + 242 to arrive at C^t, as is done (in fn. 29, p. 940), would have been to add 3648 + 242.

[4] Behrman and Birdsall calculate C^p, which they label "average annual value while in school of postponing entry into the work force to attend school" (p. 940), as the solution for C^p of

$$(a)\quad \bar{Y}\int_{S+1}^{N+S+1} e^{-rt}\,dt = C^p \int_1^S e^{-rt}\,dt,$$

But the left side of (a) is the present value of mean lifetime earnings (with schooling fixed at $\bar{S}$) if earnings were to commence one year after the last school year, rather than the present value of lifetime gross gains from schooling. The Behrman-Birdsall interpretation of C^p is therefore incorrect. In addition, the right-hand term is integrated over $S-1$ rather than S years of schooling.

of schooling is then the solution for r' of

$$(4)\quad 154(-1/r')[e^{-r'(41)}-e^{-r'}] = 1464(-1/r')(e^{-r'}-1),$$

which is $r' = 9.8$ percent.[5]

II. A Test of the Behrman-Birdsall Model

A. *Some Thoughts on Specification*

Behrman and Birdsall cite several possible sources of specification and measurement error which are relevant to the study of the schooling-earnings relation in developing countries, including omission of variables to control for individual differences in general learning environment, motivation and ability; misrepresentation of schooling quality; inappropriate regional aggregation and inadequate measures of the true cost of schooling. Most of these problems are at least partially the result of the fact that the typical data set used to study the schooling-earnings relation in a developing country is from a national source with serious limitations. Other sources of bias in this context include failure to control for cost of living differences between regions,[6] and failure to include fringe benefits in earnings.[7]

Many of these problems are reduced or eliminated if the schooling-earnings relationship is estimated using local survey data. Besides being based on a population with less variation in such factors as the general learning environment, ability, and motivation, which may obscure the relationship between returns to schooling quantity and quality, local surveys are less likely to suffer from cost-of-living-induced distortions. Moreover, some local surveys provide information not normally available in national surveys. These advantages come at a cost. Obviously, national generality is lost. Also, estimates based on local surveys probably introduce a truncation problem, since those who are educated in the survey area and subsequently migrate are not included in the sample. If successful migration is selective of those who benefit most from schooling quantity or quality, estimated rates of return will tend to underestimate true rates of return. This problem is particularly important if the survey is conducted in a relatively underdeveloped area of the country. Nonetheless, local survey data provide a good test for the robustness of the Behrman-Birdsall model.

[5]If a corrected Behrman-Birdsall "averaging procedure" is used, estimated r' is 10.2 percent. One might argue that it is the last year rather than an additional year of schooling which is relevant to the marginal analysis. Using the last year does not alter qualitative results ($r' = 9.6$ percent). Moreover, the Behrman-Birdsall result with respect to quality is based on an additional unit of quality. The estimate is, of course, also sensitive to assumptions made with respect to the social cost of an additional year of education. If, for example, it were assumed that each teacher had the 1971 national average number of student for public primary schools, and that nonsalary costs are equal to one-third of teachers' salaries, r' would be estimated as 10.2 percent by the marginal method, and 10.4 percent by the Behrman-Birdsall averaging procedure.

[6]Cost-of-living considerations are likely to be particularly important in a country such as Brazil, where large regional and rural-urban wage differences are accompanied by large differences in the cost of living. In 1969–70, the Brazilian minimum wage, which is formulated as a weighted average of survival requirements for food, housing, clothing, health and transport in 23 subregions of the country, was nominally 59 percent higher in the (most developed) state of São Paulo than in the (least developed) northeastern states. Individuals from the more developed areas tend to have more years and higher quality of schooling. Use of unadjusted data from a national census, as in the Behrman-Birdsall study, may therefore impart an upward bias to the estimated rates of return to quantity and quality of schooling.

[7]The Brazilian census provides no information with respect to fringe benefits. But these benefits are substantial to those employees who are lucky enough to possess a labor ministry card signed by the employer. Federal law mandates benefits for those individuals that include a thirteenth month's salary, paid in December, an indexed and interest earning fund (FGTS) to which employers contribute 8 percent of the employee's monthly salary (with no deduction from salary), that the employee receives upon leaving the employ of that firm or upon retirement; and social security-type benefits. There is little doubt that the probability of access to these benefits increases with quantity and quality of schooling. Failure to include fringe benefits in earnings is therefore expected to impart a downward bias on the estimated rates of return to schooling quantity and quality.

Another possible problem has as its source the measurement of labor force experience. The estimated coefficients of schooling quantity and/or quality may be sensitive to the measurement of experience. When actual labor force experience is not available, a measure of potential labor force experience is commonly used as a proxy. The conventional representation of potential experience in age minus years of schooling minus k, where k is the average age at which schooling begins. Behrman and Birdsall represent potential experience as "...the number of years since leaving school one has been 15 or older" (p. 933). The conventional representation is more consistent with the Brazilian situation. It is common to find men under the age of 15 actively involved in the labor market. In fact, the age boundary of the official Brazilian statistical category "economically active population" is 10. Using the Behrman-Birdsall measure, labor market experience of individuals with few school years (mean school years from their sample is 3) is inaccurately measured. The theoretically expected and empirically confirmed positive association between years of schooling and quality of schooling (Behrman-Birdsall, p. 934) indicates that the same inaccuracy probably occurs for individuals with low schooling quality. Finally, the conventional measure of experience is more consistent with the method used to derive the social rates of return to schooling quantity and quality, which assumes an immediate transition between schooling and labor force involvement.

These arguments provide a priori grounds for the preference of the conventional measure of potential labor force experience. On empirical grounds, the conventional measure is also preferred. The data described below were used to estimate equation (1) under the two alternative measures of potential labor force experience. A specification test of the type suggested by L. G. Godfrey (1983) was conducted on the competing models. The model with the conventional measure clearly outperforms the model with the Behrman-Birdsall measure of experience.[8]

B. *Data and Variable Definitions*

The survey upon which the following results are based was conducted in 1979 in the northeastern Brazilian metropolitan area of Fortaleza (1980 population approximately 1.5 million). The sample consists of 1,839 males aged 15 to 35 (the same demographic group chosen by Behrman and Birdsall).[9] Summary statistics with respect to the variables used to estimate a simple Mincerian specification and equation (1) from this comment are presented in Table 1. The dependent variable is equal to the logarithm of annual earnings including fringe benefits.[10] Years of schooling and years of private (primary and secondary) schooling represent schooling quantity and quality, respectively. The choice of years of private schooling is based on the fact that this is clearly a feasible higher quality alternative in Fortaleza.[11]

[8]The Godfrey (p. 362) test can be summarized as follows. Given two competing nonnested specifications (M_x and M_z), *OLS* estimates of the coefficients of one specification (say, M_x) are estimated and used to generate predicted values of the dependent variable (P_x). The variable P_x is then regressed on the independent variables of the alternative specification. The vector of least squares residuals from this regression is then used as an additional regressor (say, U) in M_x, and the augmented version of M_x is estimated with *OLS*. The regressor U is then the test variable, and its estimated coefficient (b) has a t distribution with $n-k_x-1$ degrees of freedom if the null hypothesis, $b=0$, is true. The alternative model is treated in analogous fashion. If the null is accepted, then the specification has "passed" the test. The t-statistic for the coefficient of the augmented variable in conjunction with the Behrman-Birdsall measure is -7.289, while that with the conventional measure is 1.163.

[9]Interviews of 3,240 households were administered, gathering household and individual information for 13,422 individuals over 10 years of age. Of the total, 2,977 are males between 15 and 35. Those who earned no income in the month previous to the interview are eliminated (1,072) as well as income earners for whom complete information was not available (66) to yield this sample.

[10]The fringe benefits included are FGTS and thirteenth month (see fn. 7).

[11]National statistics (*Anuario Estatistico do Brasil*, 1973, pp. 753–87) confirm the higher quality of private education. For example, the mean student-teacher ratio in public primary schools was 27, vs. 25 in private primary schools, the mean student-classroom ratio in public primary schools was 50, vs. 36 in private primary schools, the mean pass rate in public primary schools was 78 percent, vs. 89 percent in private primary schools. Similar results hold at the secondary level.

TABLE 1—MEANS, STANDARD DEVIATIONS AND BIVARIATE CORRELATIONS FOR SELECTED VARIABLES; MALES 15–35, IN FORTALEZA, 1979[a]

Variable	Mean	Standard Deviation	Bivariate Correlations			
			$\ln Y$	S	Q	E
$\ln Y$	10.601	.871	1.000			
S	7.784	4.877	.515	1.000		
Q	1.792	3.713	.397	.558	1.000	
E	11.192	6.992	.047	−.620	−.305	1.000

[a]Based on 1,839 income earners from survey supervised by author. See text for variable definitions.

Experience is defined as age minus years of schooling minus 6.

C. *Estimation Results*

Table 2 contains *OLS* estimates for a simple Mincerian model (col. 1) and a Behrman-Birdsall model (col. 2). As in the Behrman-Birdsall paper, the quality-inclusive specification is clearly preferred to a simple Mincerian specification at an empirical level.[12] In the quality-inclusive specification, estimated coefficients of quality interactive variables are of expected sign and significant.[13] Inclusion of quality variables reduces the estimated private rate of return to schooling at mean quality, although the change (from 16 to 15 percent) is much smaller than Behrman and Birdsall found using national data (from 20 to 11.7 percent). Since most existing estimates of the private rate of return to years of schooling in developing countries are from national cross sections, the conclusion that existing estimates are upward biased due to omission of quality in the specification of the schooling-earnings relation seems justified.

TABLE 2—ALTERNATIVE ESTIMATES OF LN INCOME FUNCTIONS: FORTALEZA MALES, AGES 15–35 IN 1979[a]

Right Side Variables	(1)	(2)
S	.160	.141
	(42.30)	(31.49)
E	.122	.124
	(16.45)	(16.96)
E^2	−.00186	−.00201
	(6.93)	(7.54)
$S*Q$		.00552
		(6.19)
$S*Q^2$		−.000264
		(4.20)
$\ln Y_0$	8.316	8.395
$\overline{R}^2$	.497	.512
$S.E.E.$	.618	.608
$\partial \ln Y/\partial S$ at $\overline{Q}$	.160	.150
$\partial \ln Y/\partial Q$ at $\overline{S}, \overline{Q}$	–	.0358

[a]Absolute values of t-statistics are shown in parentheses. Col. 1 contains estimates of equation (1′) from Behrman-Birdsall. Col. 2 contains estimates of equation (1) from this comment.

The principal result of estimation is a confirmation of the corrected Behrman-Birdsall conclusion with respect to rates of return to schooling quantity and quality. To calculate r' for years of schooling from the equation (1) estimate of r, the procedure of Section I is followed, with $r = .150$, $S = 7.784$, $n = 40$ and $Y = e^{10.704}$ (estimated Y for mean values of S, Q, and E), which yields $C* = 44418$, $C^s = 5948$,[14] and $r' = 13.3$ percent.[15]

[12]The F-statistic for a zero restriction test on the coefficients of the additional variables in col. 2 is 29.167. The critical F-value at the 1 percent level is 6.6.

[13]The Behrman-Birdsall quality variable controls for interregional and rural urban location of education. A specification including interaction terms consisting of quadratic forms of years of southeast schooling and years of rural schooling, as well as the variables in equation (1), was estimated for the Fortaleza sample. A zero restriction test on these four additional variables was passed at the 5 percent level. The F-statistic is 1.74 vs. a critical value of 2.37.

[14]To derive C^s, it is necessary to estimate mean teacher earnings ($\hat{Y}_T$). The estimate of $\hat{Y}_T$ used here was generated using the estimated coefficients from equation (1), under the assumptions that $S = 8.8$ (the

Calculation of the social rate of return to schooling quality at mean S and Q is based on the following assumptions: 1) the additional year of private schooling occurs in the middle of the student's schooling; 2) the impact on the income stream begins only after schooling is complete; and 3) the present value of the social cost of an additional year of private schooling is Cr$6,000 (about $200 in 1979). The first two assumptions are consistent with the Behrman-Birdsall methodology (pp. 938–40). Little hard data are available with respect to the cost of private education in Fortaleza in 1979, but the assumption of Cr$6,000 is probably conservative.[16] The social rate of return to quality of schooling is then the solution for q to

$$(5)\quad 1624(-1/q)[e^{-q(44.392)}-e^{-q(4.392)}] = 6{,}000,$$

which yields $q = 14.4$ percent.

Virtually the same relationship is thus found between the estimated social rates of return to schooling quantity and quality using local survey data as is found in the corrected version of the Behrman-Birdsall study.[17] This result is sensitive to the specification of experience.[18] Based on the findings of the empirical analysis of this comment, the corrected findings of Behrman-Birdsall, and the potential sources of bias noted in both, there is insufficient evidence to warrant a redirection of resources by the governments of developing countries toward deepening and away from widening educational investments.

III. Remaining Problems

No evidence has been found to indicate substantial inefficiency in the allocation of resources between schooling quantity and quality in Brazil. Several important issues remain with respect to both the relative and absolute social rates of return to investments in schooling quantity and quality. The truncation problem associated with the Fortaleza results is probably a source for downward bias in both estimates rates of return. The national result may contain a downward bias to the estimated social rate of return to schooling quantity due to the use of the Behrman-Birdsall specification of the experience variable.

Neither analysis deals with the fact that individuals may be engaged in full-time work while they are going to school (24 percent of the Fortaleza sample). Externalities associated with schooling quantity and quality are also ignored. The omission of treatment of

average number of school years for teachers from the Behrman-Birdsall sample); $E = 20$; and the remainder of the variables are equal to their respective mean values from the Fortaleza sample.

[15] By setting years of private primary and secondary schooling (Q) equal to zero, a social rate of return to years of public schooling can be estimated. The private rate of return to years of public schooling is 14.1 percent from equation (1). Adjustment for social cost yields an estimate of 12.4 percent for r'.

[16] Cr$6,000 corresponds approximately to annual tuition cost at private middle schools in Fortaleza during 1979. Additional fees for educational materials (of at least 50 percent of tuition cost) were also normally charged by private schools.

[17] The inclusion of variables to control for ability and general learning environment differences does not alter this result. A simple variant of the Behrman-Birdsall model was estimated in which such controls were introduced. (Results are available on request.) Estimates of the social rates of return to quantity and quality at mean S and Q are 12.4 and 12.8 percent, respectively. The social rate of return to schooling quality thus appears to be more sensitive to the inclusion of these variables.

[18] Equation (1) was estimated with the Fortaleza data set, using the Behrman-Birdsall specification of experience. Resulting estimates of the social rates of return to schooling quantity and quality are 9.1 and 14.9 percent, respectively. See the text and fn. 8 for a priori and empirical reasons for preferring the use of the more conventional specification of experience. If the estimates of social rates of return based on national data are equally sensitive to the specification of experience, it is expected that the national estimate of the social rate of return to schooling quantity would be greater than that to schooling quality. Such a result would of course be subject to the remaining potential sources of bias mentioned in this comment and in Behrman-Birdsall. The dramatic change in estimated relative social rates of return is an indication of the need for 1) caution in the specification of the experience variable, and 2) further study with respect to postschooling investments in human capital in the context of developing countries.

these two factors probably results in estimates for both rates of return in both analyses which are biased downwards.

The only empirical treatment of the general learning environment, ability and motivation in either analysis is a superficial one (see fn. 17). It may be important to control for these factors, since educational policy is implemented in a given general learning environment (in the short run). The omission of these factors probably imparts an upward bias to the estimated social rates of return (in the short run) to both quantity and quality.

Finally, both analyses use proxies for quality. The proxies almost certainly have a lower variance than the true quality of schooling. The direction of bias in the estimated coefficients of quality is unknown under these circumstances. The importance of this final factor is that the other problems imply biases in the same direction for both estimated rates of return, while this problem implies a bias which is applicable to only the estimated rate of return to schooling quality.

REFERENCES

Behrman, Jere R. and Birdsall, Nancy, "The Quality of Schooling: Quantity Alone is Misleading," *American Economic Review*, December 1983, *73*, 926–46.

Godfrey, L. G., "Testing Non-nested Models after Estimation by Instrumental Variables or Least Squares," *Econometrica*, March 1983, *51*, 355–65

Mincer, Jacob B., *Schooling, Experience and Earnings*, NBER, New York: Columbia University Press, 1974.

Brazil, Ministerio de Planejamento, *Annuario Estatistico do Brazil*, 1973, *34*, Rio de Janeiro: IBGE, 1973.

The Quality of Schooling: Reply

By JERE R. BEHRMAN AND NANCY BIRDSALL*

In our 1983 paper, we argue that the exclusion of schooling quality from standard earnings function estimates may cause some important misinterpretations of the impact of years (or quantity of schooling). As an empirical example we provide estimates for Brazilian males age 15–35 for which we report the seven following major empirical findings.

From the national estimates:

> 1. *If the true specification is quality-inclusive, the standard procedure substantially overestimates the true private rate of return to schooling.* [p. 937]
> 2. *The estimated internal social rate of return to investment in school quality is larger than it is to investment in school quantity.* [p. 938]
> 3. *There probably is an important equity-productivity tradeoff in the allocation of resources to schooling.* [p. 940]

From the regional estimates:

> [4] [*C*]*ontrol for differences in quality reduces substantially unexplained regional income differentials.* [p. 944].
> [5] *The standard specification likewise overstates urban-rural differentials for equally schooled individuals because it omits quality differentials.* [p. 944]
> [6] *The standard approach overestimates migrant-nonmigrant differentials and the probable importance of migration selectivity on unobserved individual characteristics.* [p. 945]
> [7] *Finally, the standard approach may result in overestimates of the equity-productivity tradeoff across geographical areas.* [p. 945]

In his comment, Peter Eaton agrees with the possible importance of incorporating quality into the analysis, but criticizes primarily two dimensions of our empirical analysis: 1) our calculation of the social rate of return for investments in the quantity of schooling; and 2) our definition of experience. He also provides micro estimates from the Brazilian metropolitan area of Fortaleza that, in his view, seem to reinforce the amended implications of our analysis. We discuss each of these components of Eaton's comment in turn.

I. Calculation of the Social Rate of Return to Quantity

In our article we present an estimate of the social rate of return to quantity of 6.8 percent (fn. 29), as compared to the private rate of return to quantity of 11.7 percent and a social rate of return to quality of 10.4 percent. Eaton notes correctly that there is an arithmetic error in footnote 29 (a failure to multiply a monthly earnings figure by 12 to obtain an annual figure) that causes most of the difference between our reported calculations of the private vs. the social rates of return to quantity. He also proposes a marginal instead of an average calculation, which makes little empirical difference in this instance. His amended estimate of the social rate of return to quantity is 9.8 percent,[1]

*William R. Kenan, Jr. Professor of Economics, McNeil 160/CR, University of Pennsylvania, Philadelphia, PA 19104, and Chief, Policy and Research, Population, Health and Nutrition Department, World Bank, 1818 H Street, NW, Washington, D.C. 20433, respectively. This paper was prepared as part of the World Bank Research Project 672-21, "Studies on Brazilian Distribution and Growth." We are grateful to Punam Chuhan for efficient programming. The World Bank is not responsible for any view expressed here.

[1]As indicated in Table 1, by our calculation using Eaton's approach this rate of return should be 9.5 percent, with the small difference between his and our calculation apparently due to rounding errors in his calculation of the mean ln income.

which is not substantially less than the estimate of 10.4 percent for quality.

We agree with Eaton's corrections and thank him for them. As a result, the second of our seven conclusions listed in the introduction, at least for the specifications in our article (but see below), should be amended to read: 2. *The estimated internal social rate of return to quality is at least as large as is that to investment in quantity*. Our other six major findings are not affected by this correction. In his third paragraph, Eaton suggests that our third finding, of a probable equity-productivity tradeoff, also is affected. But it is not. A sufficient basis for an equity-productivity tradeoff, if one starts from equilibrium, is the interaction between quantity and quality discussed in our article.[2]

II. Proxy for Experience

We utilized as a proxy for potential "adult" labor force experience "...the number of years since leaving school one has been 15 or older" (p. 933). Eaton argues that the conventional proxy, the number of years since leaving school, regardless of the age at leaving school, is preferable on theoretical grounds, more consistent with our derivation of social returns (which assumes an immediate transition from school to work), and preferable on empirical grounds for his Fortaleza sample.

The standard proxy is more consistent with our derivation of social returns. However, the assumption required of an immediate transition from school to work is a strong one, particularly in a setting where many children attend school for only a few years (the mean years of schooling for individuals in our sample is three), and thus may be leaving school at ages of 10 and younger. In our sample, moreover, the adult experience proxy is preferable on empirical grounds. Columns 1–4 of Table 1 give a comparison of the national estimates with the alternative experience definitions for both the standard model and our quality-inclusive model. For both the standard semilog relation and our quality-inclusive relation, the standard errors of estimate are smaller with the use of our adult potential experience measure than with the standard experience measure.

In case some readers, nevertheless, have a strong a priori preference for the standard experience measure, we briefly discuss how its use changes our conclusions. The use of standard experience results in substantially higher estimates of the rate of return to years of schooling and about the same estimates of the rate of return to schooling quality in our quality-inclusive specification. Based on these estimates, our original second result and the amended version of it suggested by Eaton are both incorrect. Instead, the implied social rate of return is much higher for quantity than for quality of schooling. This implies considerable current inefficiency in the allocation of schooling resources, with too many resources dedicated to school quality. Based on casual observation of the school system in Brazil, which—especially in the poor areas of the northeast—is characterized by high repetition and drop-out rates, the finding that too many resources are going to school quality compared with quantity is surprising. It certainly suggests that the standard definition of experience should not go unquestioned in settings where many children leave school well before the age of 15, but may or may not immediately begin full time work.

All of the other results, in any event, continue to hold for these estimates as well as for our earlier ones. On the national level, for example, the standard quality-exclusive estimate using the standard experience measure overstates the private rate of return to schooling by 8.5 percent at the point of sample means if the true model is the quality-inclusive one, as compared with an overstatement of 8.8 percent in our original comparison with the adult experience measure. Thus our first finding still holds. Likewise, due to the interaction between quality and quantity, the third finding of a productivity-equity tradeoff still holds if one starts from equilibrium (with the social rates of return to

[2] The text on p. 941 of our earlier article might suggest that the third finding also is affected, since the first point there refers to the second finding. The quantity-quality interaction is mentioned as the second point on p. 941.

TABLE 1—STANDARD AND PREFERRED QUALITY-INCLUSIVE LN INCOME FUNCTIONS WITH ADULT AND STANDARD POTENTIAL EXPERIENCE MEASURES FOR BRAZILIAN MALES, AGE 15–35 IN 1970[a]

	Standard Relation		Quality-Inclusive Relation		Extended Quality-Inclusive Relation
Right Side Variables	Adult Experience (1)	Standard Experience (2)	Adult Experience (3)	Standard Experience (4)	Adult Experience (5)
S	.205	.325	−.185	−.064	−.135
	(38.1)	(51.7)	(2.6)	(0.9)	(1.9)
$S \cdot Q$			.037	.039	−.0027
			(2.4)	(2.5)	(0.2)
$S \cdot Q^2$			−.0003	−.0005	.0040
			(0.3)	(0.6)	(4.3)
S^2					.0350
					(4.6)
S^3					.0013
					(5.6)
S^2Q					−.0057
					(9.1)
E	.303	.249	.304	.254	.303
	(25.6)	(17.9)	(26.1)	(18.6)	(26.3)
E^2	−.0090	−.0038	−.0091	−.0040	−.0091
	(15.6)	(9.2)	(16.0)	(9.9)	(16.2)
$\ln Y_0$	2.15	0.74	2.24	0.81	2.19
$\bar{R}^2$	.345	.326	.391	.350	.381
$S.E.E.$	1.444	1.464	1.415	1.438	1.403
Private Rate of Return to S[b]	20.5	32.5	11.7	24.0	9.5
Social Rate of Return to S[b]	17.1	27.0	9.5	20.2	7.5
Social Rate of Return to Q[b]	–	–	10.4	9.6	10.0

[a] S = years of schooling or schooling quantity, Q = schooling quality, E = experience, and Y_0 = the zero-schooling income level. The data, models, and most of the derivations are discussed in our 1983 article (though there is a typographical error in relation (6′) in that the penultimate, not the last term, should be divided by q). Eaton's method is used to calculate the marginal social rate of return to schooling. The rates of return all are calculated at the sample means. The absolute values of the t-statistics are shown in parentheses. Adult experience refers to potential postschooling experience above the age of 15. Standard experience refers to age minus years of schooling minus six.

[b] Shown in percent.

quality and quantity equated), though these estimates suggest that—starting from the actual location—initial broadening of schooling would be most productive and also would move Brazilian society towards more equal distribution.

However, the result regarding the relative magnitudes of the implied social rates of return to quality and to quantity of schooling is sensitive to the exact specification. To illustrate, Table 1 (col. 5) gives estimates of an extended quality-inclusive model with adult experience. We prefer this extended model over the specifications in our original paper on a priori grounds because it allows diminishing returns to schooling quality. It also is preferred empirically in the sense that the standard error is smaller for this specification than for any of the others in the table (as well as for an extended specification with the standard experience measure). With this specification, the social rate of return to schooling quality is higher than that to schooling quantity—which is consistent with the original second conclusion (and not the amended version) of our article.

III. Eaton's Estimates for Fortaleza

Eaton notes several advantages and disadvantages of using micro survey data from a small population versus our national estimates. We basically agree with the advantages that he states for each approach,[3,4] though we do not know how to trade off the relative advantages to decide which estimates are better, and do not think that Eaton knows either. We do note that Eaton's procedure assumes that all public schools have homogeneous quality and that all private schools have homogeneous quality, even though by most measures there probably is substantial variance within both types of schooling. Given such problems, it is somewhat gratifying that in general terms Eaton sees his estimates as pointing to the same findings (including the amended second one) as do our original results.

In conclusion, Eaton's comments are useful. But reflection upon them basically leads to a reaffirmation of our earlier emphasis on the importance of schooling quality and on most, if not all, of our original empirical findings. The one result that seems particularly sensitive to the exact specification is the relative magnitude of the social rates of return to quality and quantity. In our preferred quality-inclusive extended model, though not in all specifications, the social rate of return is higher for schooling quality than for schooling quantity.

[3] In fact, in our 1984 article, we have placed considerable emphasis on biases due to cost-of-living variation across space (as well as other geographical biases), which is one of the biases that Eaton emphasizes.

[4] But neither his sample nor ours controls for some of the variables that he mentions, such as "motivation and ability." For efforts to do so with special data for developing countries, see Behrman and Barbara Wolfe (1984) and Jane Armitage and Richard Sabot (1983).

REFERENCES

Armitage, Jane and Sabot, Richard, "Socioeconomic Background and the Returns to Schooling in Two Low-Income Economies," mimeo., World Bank, 1983.

Behrman, Jere R. and Birdsall, Nancy "The Quality of Schooling: Quantity Alone is Misleading," *American Economic Review*, December 1983, *73*, 928–46.

______ **and Wolfe, Barbara L.,** "The Socioeconomic Impact of Schooling in a Developing Country," *Review of Economics and Statistics*, May 1984, *66*, 296–303.

Birdsall, Nancy and Behrman, Jere R., "Does Geographical Aggregation Cause Overestimates of the Return to Schooling?," *Oxford Bulletin of Economics and Statistics*, February 1984, *46*, 55–72.

Eaton, Peter J., "The Quantity of Schooling: Comment," *American Economic Review*, December 1985, *75*, 1195–201.

The Economics of Performing Shakespeare: Comment

By EDWIN G. WEST*

In a recent article in this *Review* (1984) James Gapinski provides an interesting attempt to determine whether the benefits of public patronage going to the Royal Shakespeare Company (*RSC*) exceed the cost. The method adopted involves the measurement of the benefits from public subsidy as the resulting gain in consumer's surplus. Gapinski concludes that since the quantified benefits of real consumers' surplus (£900,204) exceed the real costs of patronage (£761,937) by the proportion of 1.18, the subsidy more than "pays" for itself and is therefore justified.

I wish to make the following points: (a) a complete analysis requires that allowance be made for the fact that real world taxation entails deadweight loss, and that the cost of it should be included accordingly; (b) plausible estimates of deadweight loss are more than enough to reduce Gapinski's benefit-cost ratio to less than unity; (c) a closer scrutiny of "patronage," that involves corporate as well as public donations, indicates that Gapinski's market demand curves for theater entertainment are underestimates since there are additional (joint) demand curves requiring to be (vertically) added; (d) before one can declare that a public subsidy is "justified," some discussion is required of the nature of those subsets of the population that consume the benefits. Since the typical theater audience is disproportionately educated and enjoys above average income, serious issues of distribution are involved that are not resolved in Gapinski's findings.

I. Deadweight Loss from Taxation

It is well known that in the real world a perfect (neutral) tax system does not exist. Imperfect tax institutions meanwhile entail deadweight loss and, following Joseph Stiglitz and Partha Dasgupta (1971), Anthony Atkinson and N.H. Stern (1974), and Edgar Browning (1976), this fact leads to a modification of Samuelson's rule for optimum public good supply. Whereas the latter requires that the sum of marginal benefits equal conventional marginal costs, the presence of deadweight loss dictates that the former should exceed the latter by some appropriate magnitude.

Dan Usher (1983) has shown that the logic of this modification to the Samuelson rule requires a fundamental reorientation of benefit-cost analysis. Because of deadweight loss from taxation, a public sector project should be undertaken, Usher argues, if and only if the ratio of benefits to cost exceeds a certain parameter S, that depends on the prevailing tax structure. He shows that

$$(1) \qquad S = 1/(1+\tau\varepsilon),$$

where τ is the tax rate and ε is the price elasticity of demand for goods. He also demonstrates that S, the private cost of public funds, is greater than 1 because the price elasticity of demand ε is negative. On this reasoning, Gapinski's benefit-cost estimate of 1.18 does not pass muster until compared with the variable S. This leads of course to the question what magnitude of S to apply in the British case.

II. Appropriate Estimates of Deadweight Loss

Browning produced an estimate of S (in the United States) of 1.16. If this were appropriate for Britain, Gapinski's benefit-cost ratio of 1.18 would only just pass Usher's test. There are reasons, however, to believe that the relevant S parameter is much greater for Britain than for the United States. First, the tax rate τ (representing the proportion of the *GNP* taxed) is larger in Britain than in

*Carleton University, Ottawa, ON, K1S5B6 Canada.
I thank Dan Usher and Halldor Palsson for helpful comments.

the United States. If we assume similar price elasticities of demand for goods (ε), then the denominator in equation (1) is increased by the larger τ for Britain. The result is a higher S factor than for the United States. Second, Usher's analysis demonstrates a wider range of influences upon S, including the extra welfare costs of tax evasion and tax collection. The latter, in fact, turn out to be quite significant in his empirical estimates. Incorporating all these considerations Usher has estimated an S factor for Canada (which is a closer comparison for Britain) of 2.19. If applied to Gapinski's case study, such a figure would unambiguously reject his empirical benefit-cost claim.

III. The Meaning of Patronage

Although not defined in his article, the term "patronage" is employed by Gapinski simply to denote subsidy from local and central government taxation. As well, he often appears to imply that the purpose of the subsidy is that of a straight gift. At the same time, and referring to William Baumol and William Bowen (1966), he argues that the *RSC* cultural experience is partly a public good (p. 465, fn. 12). To be consistent, therefore, Gapinski requires a system of vertically added demand curves. He derives his social demand curve, however, from data relating to market (ticket) information exclusively. But, if the *RSC* production is a public good, then presumably individuals other than direct users obtain some benefit. Baumol and Bowen, for instance, argue that nonusers "take pride in the international recognition conferred on our singers and the creativity of our choreographers" (p. 383). Such utility functions must be translated into pseudo demand curves which, when added (vertically) to others, and to Gapinski's market demand curve, produces the required comprehensive social demand schedule.

If significant at the margin, these public good considerations would generate higher demand curves for the Aldwych and Stratford outputs than is shown in Gapinski's Figure 1. Conceivably, too, the consumers' surplus under the social demand curve could now be so large (compared with the surplus under Gapinski's demand curve) as to generate a benefit-cost ratio that exceeds the S factor discussed above. In this case it could indeed be argued that the subsidy pays, or more than pays, for itself. Unfortunately we have no data relating to the pseudo demand curves. The problem here, of course, is the absence of an efficient demand revelation process. Without the latter we are left with simple *assertions* that Shakespearian theater production is a public good or involves significant external benefits. The assertions moreover are usually made by interested parties. To the extent that the wider public has been consulted on the subject at all it has usually been in opinion polls. Such studies in the United States indicate ambiguous public attitudes toward art and culture. The 1973 National Study by Louis Harris and Associates (1975) for the Associated Councils of the Arts, for example, found that respondents (16 years and older) had more respect for bus drivers and professional baseball players than for art critics, poets, ballet dancers, and professional actors.[1]

The term patronage strictly includes private as well as public "donors." It has been estimated for instance that in 1980–81 about £6½ million was contributed by private business to British arts projects.[2] To describe such private company funds directed to the arts as "donations," "support," or "private philanthropy," is misleading. It can be strongly argued that such funds are intended primarily for the promotion of private companies rather than support for the arts. As a vehicle for advertising and promoting goodwill, the arts appear to be seen by more and more companies as "cheaper than sports, higher profile than education and more upbeat than health and welfare..." (Sarah Iley, 1982, p. 142). A given Masterpiece Theater production for television, for instance, may be "made possible by a grant from Mobil

[1]The respondents were asked whether they had "a great deal of respect" for various occupations or professions. Heading the list with 82 percent were doctors; poets, actors, ballet dancers and art critics scored 35 percent or less.

[2]See Muriel Nissel (1983, p. 36).

Corporation" but, simultaneously, profitable advertising is "made possible" for the Mobil Corporation.

Partnerships between businesses and the arts can best be conceptualized in terms of joint supply models. In the absence of government subsidy, the price of theater tickets to consumers is the cost of the whole theater production minus what can be obtained from the purchase by business corporations (through the arts media) of advertising and goodwill. Similarly the price of the latter is the total cost of the theater production minus earned income from ticket sales. Again, for a complete analysis, we require separate and (vertically) additive demand functions. If, finally, we allow for government subsidies and assume that these correspond to purchases in the "market" for external benefits for citizens generally, then we have at least three classes of markets all of which are integrated via a grand joint supply model.[3]

IV. Benefit Distribution

Gapinski's conclusion that a public subsidy to the *RSC* theaters is justified, neglects attention to the distribution of benefits. Some observers have warned that, normatively, it is difficult to justify public funding of the arts if the audience is composed of a small, well-to-do segment of the population. International studies of audiences in fact indicate a high- or middle-income bias. Baumol and Bowen found for the United States, "the median family income among a typical audience is roughly *twice* as high as that for the total urban population" (p. 84). Over 60 percent of the audience for each art form consisted of middle-class people in the professions and all exhibited an extremely high level of education. Estimates for Australia show that the bottom one-third of households by income class contribute about 30 percent of the costs of arts subsidy and enjoy only 10 percent of the benefits, whereas the top 25 percent of households enjoy 55 percent of the benefits while contributing only 30 percent of the costs.[4] In England, the General Household Survey shows that only 15 percent of the visitors to the theater, opera, and ballet are in the semi- and unskilled manual trades.[5]

Gapinski argues that his type of benefit-cost analysis of the arts is a useful tool for policy formulation and particularly at a time when governments seek ways to cut a government deficit. Such a policy, he suggests, might be wrong because "inasmuch as decreased funding of the arts would reduce employment and output, would raise price, would destroy surplus, and would compromise a government program that has merit on economic grounds. The issue is hardly one-dimensional" (p. 465). With the last sentence one can heartily agree. But a dimension that must be included in the discussion is surely the distribution of benefits. Times of recession especially would appear to be appropriate occasions to attempt to relieve the low-income groups generally, and the unemployed, in particular, from having to contribute taxes for the supply of services used predominantly by the well-educated, the employed and the generally well to do. As for the destruction of surplus, mentioned in the above quotation, one must repeat that Gapinski has not really demonstrated its existence in the first place. To do so, as argued above, one needs elaborate information on the various components of the social demand curve (a wider concept than the market demand curve) together with an incorporation into the analysis of deadweight loss from

[3] It is for this reason that the custom (started by Baumol and Bowen) of focusing upon what specialists in cultural economics now call the "income gap" in the performing arts is not helpful. The income gap is usually described as measuring the amount of "outside support" or patronage a company needs in order to produce and present live performing arts; it represents the difference between the costs of putting on live performances and the direct earnings at those performances. In the model outlined above, however, there is strictly no income gap. There are joint and interdependent markets for services, each of which generates sufficient income to cover its costs.

[4] See C. D. Throsby and G. A. Withers (1979, p. 188).

[5] See Nissel, p. 146.

taxation, for such loss, according to current empirical studies, is far from trivial.

REFERENCES

Atkinson, A. B. and N. H. Stern, Pigou, Taxation and Public Goods," *Review of Economic Studies*, January 1974, *41*, 119–28.

Baumol, William and Bowen, William G., *Performing Arts—The Economic Dilemma*, New York: Twentieth Century Fund, 1966.

Browning, E. K., "The Marginal Cost of Public Funds," *Journal of Political Economy*, April 1976, *85*, 283–98.

Gapinski, James H., "The Economics of Performing Shakespeare," *American Economic Review*, June 1984, *74*, 458–66.

Harris, Louis ands Associates, *Americans and the Arts*, New York: Associated Councils of the Arts, 1975.

Iley, Sarah J., "Sponsorship Patterns in Canada: From Charitable Donations to Marketing Tool," paper delivered to World Congress on Sponsorship of Sports and Arts, November 1982.

Nissel, Muriel, *Facts About the Arts*, London: Policy Studies Institute, September 1983.

Stiglitz, J. E. and Dasgupta, P. S., "Differential Taxation, Public Goods, and Economic Efficiency," *Review of Economic Studies*, 1971, *38*, 151–74.

Throsby, C. D. and Withers, G. A., *The Economics of the Performing Arts*, New York: St. Martins Press, 1979.

Usher, Dan, "The Private Cost of Public Funds," working paper, Queen's University, 1983.

The Economics of Performing Shakespeare: Reply

By James H. Gapinski*

> King. Have you heard the argument?
> Is there no offence in't?
> Hamlet. No, no; they do but jest,
> poison in jest; no offence i' th' world.
>
> *Hamlet*

The comment by Edwin West addresses my 1984 benefit-cost analysis and, more specifically, its conclusion that the subsidy received by the Royal Shakespeare Company (*RSC*) is justified. West maintains that a broader inquiry may reverse that judgment, and he raises three main points for discussion: deadweight loss, pseudo demands, and benefit distribution.

I shall consider these matters seriatim. First, however, it should be recalled that the original analysis took a narrow stance deliberately and that footnote 14 conveyed the message. Because of difficulty in determining how much patronage went to Aldwych and Stratford activities alone, *RSC* patronage obtained from all sources, public and private, was treated as if it applied only to those two centers. But, besides performing there, the *RSC*, over the financial years 1968–69 to 1977–78, worked The Warehouse, The Other Place, Theatregoround and other domestic tours, overseas proscenia, and television inter alia. Therefore assigning total patronage exclusively to the centers imparted a bias that favored the case against the subsidy. Nevertheless, the investigation did omit deadweight loss.

Prompted by West's remarks on the subject, I resurrected the data to try to obtain a rough but reasonable estimate of Aldwych and Stratford patronage. In the exercise, patronage was distributed conceptually across all *RSC* activities on the premise that each activity shared the total in the same proportion that it shared total expenses.[1] No attempt was made to separate private gifts from public ones. Table 1, which presents the pertinent details, indicates that nominal patronage going exclusively to the centers averaged £513,778, about 85 percent of the total. With the mean of the Retail Price Index amounting to .801, this figure becomes £641,421 in real terms. Benefit, the increase in real consumers' surplus at the centers, still registers £900,204 enabling the benefit-cost ratio to rise from 1.18 to 1.40.[2]

These calculations ignore an important side effect of the subsidy; namely, the additional tax collections coming from both increased ticket sales and increased labor income. As the earlier Table 3 indicated, a profit-maximizing Aldwych sells 16,528 tickets at a real price of £3.78 including the value-added tax (*VAT*). A *VAT* rate of 8 percent, not inappropriate for the period in focus, therefore means a pretax real price of £3.50 and a real *VAT* levy of £0.28 per ticket for a total of £4,628. Under actual (subsidized) conditions the Aldwych sells 234,045 tickets at a real tax-inclusive price of £1.77 and pays a real *VAT* of £30,426, an increase of £25,798 over the profit maximizer. Similarly, actual Stratford generates £9,153 more in real *VAT* than does a profit-maximizing Stratford bringing the combined *VAT* increase to £34,951.[3] Table 3 likewise showed that patronage expands employment by 461,448 man-hours at

*Florida State University, Tallahassee, FL 32306.

[1] This premise follows the allocation rule referenced by Steve Barsby (1972, p. 14).

[2] If Aldwych and Stratford patronage were computed as the mean of reals, it would read £577,991, and the benefit-cost ratio would be 1.56, not 1.40. The method used in the text here duplicates that found in the text of the original paper.

[3] Tax revenues may suffer as consumers are lured to the *RSC* from other pursuits. However, tax losses from reduced endeavors outside the performing arts may be more than recouped by collections derived from the amenity package that, according to William Baumol and William Bowen (1966, pp. 261–63; 498) and to Harry Kelejian and William Lawrence (1980, p. 335), often accompanies a night at a live performance. Due to the absence of data, such ancillary minuses and pluses are presumed to exactly cancel. This presumption holds throughout.

TABLE 1—BUDGETARY CHARACTERISTICS OF THE *RSC*[a]

Financial Year	Nominal Expenses: Total	Nominal Expenses: Aldwych and Stratford	Expense Ratio[b]	Nominal Patronage: Total	Nominal Patronage: Aldwych and Stratford
1968–69	1,149,917	904,544	.787	229,650	180,735
1969–70	1,129,772	975,955	.364	236,450	204,293
1970–71	1,231,480	1,011,058	.321	280,670	230,430
1971–72	1,425,203	1,173,398	.323	303,100	249,451
1972–73	1,534,515	1,362,369	.388	430,124	381,950
1973–74	1,745,297	1,487,179	.852	409,387	348,798
1974–75	2,000,781	1,829,471	.914	694,011	634,326
1975–76	2,409,202	2,004,028	.832	792,510	659,368
1976–77	3,174,084	2,576,736	.812	1,231,919	1,000,318
1977–78	3,825,442	3,186,246	.833	1,498,335	1,248,113
Mean	1,962,569	1,651,098	.843	610,616	513,778

Sources: RSC annual reports, *Report of the Council*, Council of the Royal Shakespeare Theatre and findings by Trevor Gambling and Gordon Andrews (1982).

[a] Nominal figures are denominated in unit pounds.

[b] Expense ratio equals nominal expenses for Aldwych and Stratford divided by total nominal expenses.

the Aldwych and by 345,640 hours at the Stratford. With the before-tax real wage averaging £1.27 in the South East and £1.29 in the West Midlands for the period, the employment gains translate into extra real income-tax receipts of £361,170, behind which tally stands the marginal tax rate of 35 percent reported by Joseph Pechman (1980, p. 245) for the U.K. experience.

The real *VAT* and income-tax increments sum to £396,121. They partially offset the *RSC* subsidy, and they boost the benefit-cost ratio to 3.67, a value well above the 2.19 guide cited by West for a comprehensive situation that covers deadweight loss and other concerns.[4] If one cares to continue the drill by including the benefits associated with the pseudo demands mentioned by West, then the arithmetic becomes even more compelling in favor of the subsidy.

On the distribution of benefits, the prior effort ignored it in keeping with a strong tradition which dates back at least to the mid-1930's and which, the revisionists notwithstanding, has been sustained by the likes of Richard Musgrave (1969, pp. 803–04), Arnold Harberger (1971, pp. 785; 795), and E. J. Mishan (1982, pp. 30–31; 42–43). Perhaps West's quarrel should be directed more at the U.K. tax-transfer system in general than at the *RSC* subsidy in particular. Furthermore, his description of the *RSC* audience as a small and affluent segment of the population misses the community involvement that typified the Company during the period. In describing the *RSC*'s Theatregoround, begun in 1965, David Addenbrooke observes that "two or three times a week, six actors, using a portable stage and a handful of props, staged performances in schools, youth clubs, canteens, evening institutes, Town Halls and housing estates—in fact anywhere that a group of people could gather to form an audience" (1974, p. 68). Although Theatregoround vanished in 1971, it reappeared in spirit a few years later as Sally Beauman (1982, pp. 311; 344–45) notes.

Numbers entering benefit-cost computations ordinarily entail doses of arbitrariness. Dan Usher (1983, pp. 40–47), whose 2.19 result becomes the gauge for West, confirms this point himself and argues that 2.19 should be regarded as mnemonic rather than as law.

[4] Precedents for cost offsets in applied studies are given by Barsby (pp. 122; 127) and by the U.S. Congressional Budget Office (1976, pp. vi; 36–41). Richard Musgrave and Peggy Musgrave (1973, p. 179) present the textbook view.

Indeed the figures leading to the 3.67 ratio have their share—or possibly more than their share—of arbitrariness, and therefore quibbles either way are almost inevitable. Yet given the magnitudes involved, there seems to be little reason for rejecting the original conclusion that the *RSC* subsidy is justified. It is, despite the comment by West.

In the present context, Hamlet should be interpreted literally.

REFERENCES

Addenbrooke, David, *The Royal Shakespeare Company: The Peter Hall Years*, London: William Kimber, 1974.

Barsby, Steve L., *Cost-Benefit Analysis and Manpower Programs*, Lexington: Lexington Books, 1972.

Baumol, William J. and Bowen, William G., *Performing Arts—The Economic Dilemma*, New York: Twentieth Century Fund, 1966.

Beauman, Sally, *The Royal Shakespeare Company: A History of Ten Decades*, Oxford: Oxford University Press, 1982.

Gambling, Trevor and Andrews, Gordon, "An Analysis of the Personnel Costs of a Major Theatrical Company: 1968–78," unpublished paper, University of Birmingham, England, 1982.

Gapinski, James H., "The Economics of Performing Shakespeare," *American Economic Review*, June 1984, *74*, 458–66.

Harberger, Arnold C., "Three Basic Postulates for Applied Welfare Economics: An Interpretive Essay," *Journal of Economic Literature*, September 1971, *9*, 785–97.

Kelejian, Harry H. and Lawrence, William J., "Estimating the Demand for Broadway Theater: A Preliminary Inquiry," in William S. Hendon et al., eds., *Economic Policy for the Arts*, Cambridge: Abt Books, 1980, 333–46.

Mishan, E. J., "The New Controversy about the Rationale of Economic Evaluation," *Journal of Economic Issues*, March 1982, *16*, 29–47.

Musgrave, Richard A., "Cost-Benefit Analysis and the Theory of Public Finance," *Journal of Economic Literature*, September 1969, *7*, 797–806.

______ **and Musgrave, Peggy B.,** *Public Finance in Theory and Practice*, New York: McGraw-Hill, 1973.

Pechman, Joseph A., "Taxation," in Richard E. Caves and Lawrence B. Krause, eds., *Britain's Economic Performance*, Washington: The Brookings Institution, 1980, 199–259.

Usher, Dan, "The Private Cost of Public Funds," unpublished paper, Queen's University, 1983.

West, Edwin G., "The Economics of Performing Shakespeare: Comment," *American Economic Review*, December 1985, *75*, 1206–09.

Council of the Royal Shakespeare Theatre, *Report of the Council*, Stratford-upon-Avon, England, various years.

U.S. Congressional Budget Office, "An Economic Analysis of the Full Employment and Balanced Growth Act of 1976," unpublished paper, Washington, 1976.

Can Unemployment Be Involuntary?: Comment

By LORNE CARMICHAEL*

The last few years have seen the publication of several papers on "efficiency wage" models of unemployment.[1] The chief advantage of this approach seems to be its prediction that "genuine involuntary unemployment"[2] can exist in a labor market with no artificial constraints on wage flexibility, and where all agents are rational income (or utility of income) maximizers. The result depends on the existence of a positive relationship of sufficient strength between the productivity of a worker and the wage he is paid. With this assumption the firm's profit-maximizing wage will be independent of labor supply conditions, and unemployed workers will find employers unwilling to hire them even if they offer to work at lower wages.

One of the major objections to the versions of this model in which higher wages reduce the worker's incentives to shirk, or to quit, is the fact that unemployed workers should offer to pay "entrance fees" or bonds to work at these firms. Since the entrance fee is a sunk cost once the worker actually starts his job, it will not affect his incentives to work hard or quit prematurely. The entrance fee will reduce the expected value of the firm's offer down to that of the alternative, and the involuntary nature of the unemployment disappears. There have been two responses to this argument. The first is that imperfect capital markets may prevent workers from paying a sufficiently large fee, and the second is that the firm may simply take the bond, claim that the worker has shirked, and fire him.

There are now a handful of papers that suggest that an exact solution to the moral hazard problem can be achieved.[3] The purpose of this comment is to show that the presence of an imperfect capital market, or an imperfect mechanism (such as reputation) for preventing moral hazard is *not* sufficient to prevent the emergence of bonding and the elimination of unemployment as a mechanism for enforcing worker discipline. As a consequence, the involuntary nature of the unemployment in these efficiency wage models must depend entirely on an artificial (although perhaps quite realistic) constraint such as minimum wage laws for new workers.

The argument is extremely simple. Consider first the case where firm moral hazard is not a problem, but where capital markets are absent. Consider an unemployed worker in the steady-state Shapiro-Stiglitz labor market. We assume his utility function is intertemporally separable, so that the discounted utility of staying unemployed for one more period is given by

$$V_u = U_t(u_t) + \delta V_u, \tag{1}$$

where u_t is unemployment benefits in period t, and δ is the discount factor. If this worker gets a job offer in period t to work in period

*Queen's University, Kingston, ON K7L 3N6 Canada.

[1] See, for example, James Malcolmson (1981), Steven Salop (1979), Andrew Weiss (1980), Carl Shapiro and Joseph Stiglitz (1984), and the survey by Janet Yellen (1984).

[2] The wording of the definition of this term varies across papers, but the basic idea is that an unemployed worker's reservation wage is strictly below the wage offered by firms to those new workers that they hire, but for some reason there are no vacancies.

[3] For example, in a world with perfect capital markets the workers could pay their bonds into a pension fund with the following property. If the worker stays, he gets his bond and more back as a pension. If he leaves he loses his bond, but the firm does not get the money (as it would under the usual defined benefit plans). Rather, the other workers who do stay with the firm would find their pensions increased. The firm is therefore indifferent to the size of the bond and the moral hazard problem goes away. This basic idea can be found in papers by Sudipto Bhattacharya (1984), my 1980 dissertation and 1983 article, and Malcolmson (1984).

$t+1$ at the wage w, his discounted utility is

$$V_t = U_t[u_t - c] + \delta V_w,$$

where c is the entrance fee, V_t the value in period t of the offer, and V_w the value of being employed at the wage w. Consistent with the incentive story we have $V_w > V_u$. In the absence of artificial wage constraints and firm moral hazard, it is obvious that in a competitive environment the entrance fee for such a job will be bid up until $V_t = V_u$. The only requirement is that, as the worker approaches starvation in the current period, his utility can be made arbitrarily low relative to the utility of working.

The point is so obvious that it barely needs restating. When capital markets are imperfect, an efficiency wage firm cannot charge a fee that is large enough to reduce the expected value of a worker's lifetime wages down to that in his alternative. Fewer workers will therefore be hired than in a world of perfect capital markets, and the unemployment rate will be higher. However, the firm can certainly charge a fee sufficient to make the expected *utility* of its offer equal to that of the alternative. The *involuntary* nature of the unemployment disappears.

Explicit consideration of the moral hazard problem under an imperfect enforcement mechanism such as the firm's reputation will also lead to the same result. This time as the entrance fee grows, the probability of a fire rises. So long as this relationship is continuous, then it is again obvious that an equilibrium with the expected income of the alternative equal to the expected income of the firm's offer will obtain, this time since the probability of a fire has gone up. Unemployment is again higher than in a world with no moral hazard, but it is not involuntary since the expected utility of an offer to an unemployed worker equals that of his alternative.

There are, of course, other constraints that can be put on first-period wages to retain the involuntary unemployment. Minimum wage laws are perhaps the best example. In this sense, perhaps, the lesson from these models is that minimum wage laws can have important effects on markets other than those for unskilled and teenage workers.[4] However, the involuntary nature of the unemployment will ultimately depend on an artificial wage floor just as it does in the naive Keynesian models.

[4] The old-fashioned system of apprenticeship in skilled trades in some cases required workers to pay for their training—not unreasonable since they were taking up other workers' time, using up materials, and their output was not of sufficient quality to be sold. See Masanori Hashimoto (1982) for an analysis of the effects of minimum wages on training.

REFERENCES

Bhattacharya, Sudipto, "Tournaments and Incentives; Heterogeneity and Essentiality," Finance Working Paper No. F-4, Berkeley Business School, March 1984.

Carmichael, H. Lorne, "Implicit Contracting and Seniority Rules," unpublished doctoral dissertation, Stanford University, December 1980.

______, "Firm Specific Human Capital and Promotion Ladders," *Bell Journal of Economics*, Spring 1983, *14*, 251–58.

Hashimoto, Masanori, "Minimum Wage Effects on Training on the Job," *American Economic Review*, December 1982, *72*, 1070–87.

Malcolmson, James, "Unemployment and the Efficiency Wage Hypothesis," *Economic Journal*, December 1981, *91*, 848–66.

______, "Work Incentives, Hierarchy, and Internal Labour Markets," *Journal of Political Economy*, June 1984, *92*, 486–507.

Salop, Steven, "A Model of the Natural Rate of Unemployment," *American Economic Review*, March 1979, *69*, 117–25.

Shapiro, Carl and Stiglitz, Joseph E., "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, *74*, 433–44.

Weiss, Andrew, "Job Queues and Layoffs in Labour Markets with Flexible Wages," *Journal of Political Economy*, June 1980, *88*, 526–38.

Yellen, Janet, "Efficiency Wage Models of Unemployment," *American Economic Review Proceedings*, May 1984, *74*, 200–05.

Can Unemployment Be Involuntary?: Reply

By CARL SHAPIRO AND JOSEPH E. STIGLITZ*

The central message of the efficiency wage models[1] is that firms may not lower their wages, even in the face of unemployment, if net productivity depends on wages. Lower wages could increase rather than decrease a firm's net labor costs. As a result, wage rigidities and unemployment arise in efficiency wage models.

Since these models were first promulgated, labor market economists have questioned whether one could avoid involuntary unemployment with more complex contractual arrangements. In the context of the labor turnover model, the question has been asked, "Could not workers themselves pay for the turnover costs?" Lorne Carmichael, in his comment on our 1984 paper, poses the question, Could not individuals purchase their jobs? If they could, would not the "price of a job" clear the "market for jobs," in the sense that a marginal individual would be indifferent between remaining unemployed and purchasing a job?

At one level, the issue here is just a matter of semantics. It is little different from the old story that so long as there is a competitive labor market anywhere in the economy (grape picking in California) all unemployment must be voluntary, since any individual could have moved to California ("purchased a job"). To us, the fact that during the Great Depression 20 or 25 percent of the labor force in Chicago, workers who were once gainfully employed, were sitting idle at home, willing to work at the going wage in Chicago, suggests a massive market failure, regardless of whether one says that, because of their decision not to migrate to California, they were voluntarily unemployed. As Carmichael points out in his comment, even if the market for job openings operated smoothly, there still might be unemployment; only the label we attach to that unemployment would change.

Our concern in developing the efficiency wage model has been to help explain the level of unemployment and its fluctuations, and to ascertain whether, under the stipulated circumstances, the market outcome is efficient. Our paper shows that the market outcome is generally *not* efficient, regardless of how one labels the unemployment. This would be true whether firms could charge applicants for jobs or not. There are grounds for government intervention to alter the unemployment rate. At the same time, our analysis provides an important caution to government policy: some policies, such as mandatory unemployment insurance, aimed at alleviating the problems of the unemployed, may actually lower welfare and increase the level of unemployment. These results are quite independent of whether the unemployment is labelled voluntary or not.

The efficiency wage theory also provides an explanation of wage dynamics: it explains why one firm may be slow to lower its wages, until other firms do so. In fact, in the labor turnover version (Stiglitz, 1974) there may exist multiple equilibria. There may exist a high-wage equilibrium with a high level of unemployment and a low-wage equilibrium with a low level of unemployment. This result is also independent of the feasibility of workers buying their jobs.

There is a second level at which the possibility of entrance fees for jobs is not of much concern: the fact of the matter is, in modern economies individuals do not for the most part purchase jobs. Workers also do not bear

*Woodrow Wilson School of Public and International Affairs and Department of Economics, respectively, Princeton University, Princeton, NJ 08544.

[1] These models have a long history. References to the importance of the connection between worker quality and productivity can be found in Marshall (1928) and Weber (1925). The more recent interest in development economics dates to the work of J. Leibenstein (1957), James Mirrlees (1975), and Stiglitz (1976). In macroeconomics, early contributions include those of Stiglitz (1974), Steven Salop (1979), Andrew Weiss (1980), and G. A. Calvo (1979). For a more extensive survey, see Stiglitz (1984).

all of the training and turnover costs themselves, so firms are concerned with the turnover of their labor force. Therefore, the efficiency wage theory is directly relevant.

Still, the *possibility* that individuals might purchase jobs or pay for their training costs poses two difficult questions to which we have given some thought over the past few years. First, *why* do we not see job purchases more frequently? Second, are there conditions under which there is unemployment (job rationing) in the narrow sense of the word despite the possibility of entrance fees?

The answer to the second question is "yes." Indeed, the original nutritional wage productivity models are precisely of that form: forcing individuals to accumulate capital to finance a job purchase may have deleterious effects on their on-the-job productivity. But even in more developed economies, the answer may be yes. Among individuals with capital, it may be the less able that apply for employment at any given "job price;" the more able can do better by being self-employed. Thus, as the price of a job is increased, the mix of applicants may change adversely. These quality-efficiency wage effects interact with the incentive-efficiency wage effects in such a way as to result in unemployment.

Efficiency wage theory provides further insights into why firms may not be able to charge the market-clearing price for jobs (besides the adverse selection effects that arise when the price of a job is increased). There may also be important incentive effects: workers' incentives may be adversely affected if they feel that they have been unfairly treated. In terms of our model, effort e may decrease, and the propensity to quit, b, may increase with increases in the entrance fee at a particular job. These changes will necessitate a higher wage w, in order to induce workers not to shirk. The firm's total wage costs may actually increase as a result of increasing the price charged for the job.

But this is not the only argument against the use of fees for purchasing jobs. Another important theoretical argument against entrance fees is the double moral hazard problem: individuals will be concerned about putting money up front, less the firm take their money and either fire them or make their job so unpleasant as to induce them to quit. Though there are ways of mitigating this firm moral hazard problem, such as relying on firms' reputations, none are fully satisfactory. Even small entrance fees can reduce (discontinuously) the demand for jobs, since firms could simply go into the business of collecting entrance fees and then firing workers. In our 1984 paper, we discussed bonding at some length; entrance fees are a particular way in which employees might post bonds.

Firm moral hazard problems, and the incentive and adverse selection effects of charging entrance fees for jobs, are each sufficient to explain why job purchases are not more prevalent, and why, even where they might exist, the market for jobs may not function well.

Carmichael is right, however, in pointing out that other arguments sometimes put forward against the efficacy of job purchases cannot explain involuntary unemployment (in the particular sense in which he uses the term). Thus, limitations on individuals' capital may explain why the price for jobs is low, but so long as individuals have some capital, the marginal individual will be indifferent between remaining unemployed and buying a job.

In other contexts, R. Arnott and Stiglitz (1985) have shown how risk aversion implies that workers will not pay all of their specific training costs. Similarly, risk aversion explains why individuals might be loathe to pay for a job, since there is some chance that they will be ill-matched for the job (or otherwise have to leave the job), in which case the money they have put up front will be lost. But again, risk aversion will explain why the price for obtaining a job will be low, not the persistence of involuntary unemployment.

The fact that these particular arguments —imperfect capital markets and risk aversion—cannot explain involuntary unemployment does not, in any way, reduce the force of our earlier arguments (based on efficiency wage and moral hazard considerations) for why jobs are not typically purchased, and for why, even when they are, the market for jobs does not clear.

To us, involuntary unemployment is a real and important phenonomenon with grave social consequences that needs to be explained and understood. No single, simple model will provide the explanation. But simple models can provide us insights into its various aspects and facets. The efficiency wage models (either the incentive version, as we presented it, the quality version, or the turnover version) provide us with a simple, basic insight: firms may not lower wages when doing so lowers workers' net productivity. And the effect of any wage reduction on net productivity at a given firm generally depends upon the wages being paid by other firms.

Unemployment was a problem before minimum wage legislation was passed. Such legislation may well have exacerbated the unemployment problem. But to suggest, as Carmichael does, that unemployment problems would disappear in the absence of minimum wages or other governmental interferences with the market is to fly in the face of history and common sense.

REFERENCES

Arnott, R. and J. Stiglitz, "Labor Turnover, Wage Structures and Moral Hazard," *Journal of Labor Economics*, forthcoming, 1985.

Calvo, G. A., "Quasi-Walrasian Theories of Unemployment," *American Economic Review Proceedings*, May 1979, *69*, 102–07.

Carmichael, L., "Equilibrium Unemployment: Comment," *American Economic Review*, December 1985, *75*, 1213–14.

Leibenstein, J., *Economic Backwardness and Economic Growth*, New York: Wiley & Sons, 1957.

Marshall, A., *The Principles of Economics*, London: MacMillan, 1928.

Mirrlees, J., "A Pure Theory of Underdeveloped Economies," in L.A. Reynolds, ed., *Agriculture in Development Theory*, New Haven: Yale University Press, 1975.

Salop, S., "A Model of the Natural Rate of Unemployment," *American Economic Review*, March 1979, *69*, 117–25.

Shapiro, C., and Stiglitz, J., "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, *74*, 433–44.

Stiglitz, J., "Equilibrium Wage Distributions," IMSSS Technical Report No. 154, Stanford University, 1974 (*Economic Journal*, forthcoming).

______, "The Efficiency Wage Hypothesis, Surplus Labour, and the Distribution of Income in L.D.C.'s," *Oxford Economic Papers*, July 1976, *28*, 185–207.

______, "Theories of Wage Rigidity," discussion paper, National Bureau of Economic Research, 1984.

Weber, M., *The Protestant Ethic and the Spirit of Capitalism*, New York: Scribner, 1925.

Weiss, A., "Job Queues and Layoffs in Labor Markets with Flexible Wages," *Journal of Political Economy*," June 1980, *88*, 526–38.

NOTES

The ninety-eighth annual meeting of the American Economic Association will be held in New York, New York, December 28–30, 1985.

The Professional Placement Service will be located at the Sheraton Centre Hotel, and will be open from 10:00 P.M. to 5:00 P.M., December 27; 9:00 A.M. to 5:00 P.M., December 28–29; and 9:00 A.M. to 12:00 noon, December 30.

Members wishing to give papers or make suggestions for the program for the meetings to be held in New Orleans, LA, December 28–30, 1986, are invited to send their ideas to Professor Gary S. Becker, Department of Economics, 1126 East 59th Street, University of Chicago, Chicago, IL 60637. As in the past, the sessions sponsored by the American Economic Association will be of two types: invited papers and contributed papers. Most of the invited papers will be published in the *Papers and Proceedings* issue to appear May 1987; the contributed papers will not be so published. Suggestions of topics or proposals for papers for the invited sessions should be submitted no later than February 1, 1986. Econometric studies or highly mathematical papers are not appropriate for the sessions sponsored by the American Economic Association. However, members wishing to present such papers may submit their proposals or abstracts to the Econometric Society which meets with the American Economic Association, and normally schedules a number of sessions involving such papers.

Announcing the creation of The European Economic Association (EEA): the aims of the EEA are to contribute to the development of economics as a science in Europe, to improve communication between teachers, researchers, and students in economics in the different European countries, and to develop and sponsor cooperation between universities and research institutions in Europe. Officers for 1985–86 are: J. Drèze, President; J. Kornai, President-Elect; A. Atkinson, Vice-President; L. Phlips, Secretary; P. Balestra, Treasurer. Its council is composed of some thirty economists from all over Europe.

As of January 1986, the *European Economic Review* will be the EEA's official journal, edited by H. Glejser, P. Neary, A. Sandmo, and J. Waelbroeck (Chief Editor). It will publish a special *Papers and Proceedings* issue devoted to its annual congresses. The first congress will be held in Vienna on August 29–31, 1986. Members will receive the *European Economic Review* (and the *Journal of Economic Literature* if desired). The regular annual membership fee is £30; student membership is £15. The *JEL* can be obtained at an additional £20. Fees are to be paid to the EEA, c/o Tieto Ltd, Bank House, 8a Hill Road, Clevedon, Avon, England BS21 7HH.

The Faculty Exchange Center, a nonprofit, faculty-administered program, helps arrange teaching exchanges on the college/university level, and house exchanges for study and travel for teachers and administrators at all levels of the education profession. The current roster of members will be made available to new members upon registration. For more information, send a stamped, self-addressed envelope to Faculty Exchange Center, 952 Virginia Avenue, Lancaster, PA 17603.

The annual Conference of the Association of Private Enterprise Education will be held at the La Mansion Del Rio in San Antonio, Texas, April 6–9, 1986. Emphasis is on private enterprise and entrepreneurship. Business practitioners and academicians are invited to attend. For further information, contact Dr. Calvin Kent, Director, Center for Private Enterprise, 215 Hankamer School of Business, Baylor University, Waco, TX 76798.

The annual Presidential Conference, "A Retrospective on Lyndon Baines Johnson...A Texan in Washington," will be held April 10–12, 1986. For more information, contact Natalie Datlof and Alexej Ugrinsky, Conference Coordinators, Hofstra University Cultural Center, Hofstra University, Hempstead, NY 11550 (telephone 516 + 560–5669/5670).

The Southern Regional Science Association will hold its twenty-third annual meeting in New Orleans, LA, March 6–8, 1986.

The National Women's Studies Association will hold its annual convention June 11–15, 1986, at the University of Illinois at Urbana-Champaign. The theme is "Women Working for Change: Health, Cultures, and Societies." For further information, contact Jeann Rice and Paula Gray, Coordinators, NWSA, University of Illinois, 304 Stiven House, 708 South Mathews, Urbana, IL 61801.

The University of Florida's Women in Agriculture Program announces a Conference on Gender Issues in Farming Systems Research and Extension, to be held February 26–March 1, 1986. Contact Dr. Marianne Schmink, WIA Co-Director, Center for Latin American Studies, University of Florida, Gainesville, FL 32611, for further information.

A conference, "Mergers in Health Care: The Performance of Multi-Institutional Organizations," will be held at the Campus of Project HOPE, Millwood, Virginia, April 10–11, 1986. It is supported by the Project HOPE, the Medical College of Virginia, VCU, and SRI International. For further information, contact G. R. Wilensky, Vice President for Domestic Affairs. The Project HOPE Health Sciences Education Center, Millwood, VA 22646 (telephone 703 + 837-2100).

Harvard Law School offers four or five fellowships for the academic year 1986–87 to college and university teachers of social sciences or humanities to study the fundamental techniques, concepts, and aims of law. Fellowship holders will presumably take at least two first-year courses in law, in addition to more advanced courses, and will participate in a joint seminar. The year of study will not count toward a degree. The fellowship grant covers tuition and health fee only, plus provision of office space. Applications should include a resume, a statement explaining what the applicant hopes to achieve through the study, and two letters of recommendation (mailed directly to the Chair from the referees). There is no special application form. Contact the Chair, Committee on Liberal Arts Fellowships, Harvard Law School, Cambridge, MA 02138. The deadline is January 15, 1986.

The Institute of American Cultures at UCLA offers graduate and post-doctoral fellowships to support study of Afro-Americans, Asian Americans, Chicanos, or American Indians. The stipend for graduate fellowships is $5,000 per year plus registration fees and out-of-state tuition if applicable. Postdoctoral fellowships range from $20,000 to $25,000 per year. The deadline is December 31, 1985. For further information and applications, contact the Fellowship Director of the following appropriate ethnic center at UCLA, Campbell Hall, Los Angeles, CA 90024: Center for Afro-American Studies, Rm 3111; Asian American Studies Center, Rm 3232; Chicano Studies Research Center, Rm 3121; American Indian Studies Center, Rm 3220.

The Emory University Law and Economics Center will sponsor the L + EC Law Institute for Economists, May 18–30, 1986, in Atlanta, Georgia. It is a program in law, not economics. All participants receive private rooms, most meals, books and tuition. Application requires a current curriculum vita and a statement that you plan to attend the entire program if accepted. Contact Marc Hoberman, Program Administrator, Emory University, L + E Center, Atlanta, GA 30322 (telephone 404 + 727–5771). Deadline is January 24, 1986.

American Statistical Association/National Science Foundation/Census Bureau Research Fellowships and Associateships beginning around September 1, 1986; positions at Census Bureau for one year or shorter period. General areas for research are social and demographic studies, economic measurement and analysis, and statistical methodology and computing. Requirements: for Fellows, Ph.D. and research record in relevant field (Economics, Demography, Sociology, Statistics); for Associates, at least two years graduate study in relevant field. Salaries are commensurate with qualifications and experience; also, fringe benefits and a travel allowance are provided. Apply by January 1, 1986 for Fellows; February 15, 1986 for Associates. For information on specific research topics and on how to apply, contact Dr. Peter Zadrozny, Center for Economic Studies, Room 3442–3, Bureau of the Census, Washington, D.C. 20233 (telephone 301 + 763–2490).

Call for Papers: The International Conference on Input-Output Techniques will be held at Hokkaido University, Sapporo, Japan, from July 28 to August 2, 1986. The conference is sponsored jointly by the United Nations Industrial Development Organization and the Sapporo Local Organizing Committee in cooperation with the government of Japan. To contribute or to attend, contact Ms. Anna Gelei, UNIDO, P.O. Box 300, A–1400 Vienna, Austria.

Call for Papers: The annual Conference of Economists of the Economic Society of Australia will be held August 25–29, 1986 at Monash University, Melbourne. Topics include tax reform and public sector policies, labor market, exchange rates, and commercial policies. Submit two copies of paper plus abstract with *JEL* classification by March 30, 1986, to Dr. D. K. Fausten, Organizing Secretary, Department of Economics, Monash University, Clayton, Victoria 3168, Australia (telephone 03 + 541–2345).

Call for Papers: The annual meeting of the History of Economics Society will be held at the City University in New York City, May 1986. To give a paper, act as discussant, or chair a session, contact President-Elect, Abraham Hirsch, Department of Economics, Brooklyn College, Brooklyn, NY 11210.

Call for Papers: The Eastern Economic Association meeting dedicated to honoring the 50th anniversary of Keynes' *General Theory* will be held at the Franklin Plaza Hotel, Philadelphia, April 10–12, 1986. Papers (in theoretical or applied areas of economics), program participants, and program suggestions are solicited. To

present a paper, submit two copies of a one-page abstract with *JEL* classification number and a $15 submission fee by December 15, 1985. Send all inquiries and submissions to the Eastern Economic Association, Department of Economics, U-63, University of Connecticut, Storrs, CT 06268.

Call for Papers: A symposium on the 50th anniversary of the Spanish Civil War, "World War II—Prelude," will be held June 12–13, 1986 at Siena College. Papers and proposals are invited. Contact Program Committee (World War II): Professor Thomas Kelly, History Department, Siena College, Loudonville, NY 12211.

Call for Papers: The Southern Economic Association will hold its annual meeting at the Fairmont Hotel, New Orleans, LA, November 23–25, 1986. Submit two copies of a one-page abstract and cover sheet, with $10 submission fee by February 1, 1986. For details, see the Call for Papers in the *Southern Economic Journal*, October 1985, or write Dr. Joseph M. Jadlow, College of Business Administration, Oklahoma State University, Stillwater, OK 74078.

Call for Papers: An International Conference on Economics and Psychology will be held in Haifa, July 9–11, 1986. It is sponsored by the Society for the Advancement of Behavioral Economics and the International Association for Advancement of Research in Economic Psychology. Papers that integrate the models and methods of economics and psychology are sought, and should be sent to Professor Shlomo Maital, Faculty of Industrial Engineering and Management, Technion, Haifa, Israel 32000.

Call for Papers: The Democratic Policy Commission was formed in May 1985 to examine problems and opportunities that will face the American people for the balance of the decade and the century. The Commission is accepting papers on public issues (domestic, international, and institutional) for consideration in the development of its final report in April 1986. The papers should address the likely consequences of exisiting public policy issues, projected trends, changing values, new ideas, and/or alternative policies that should be considered by the Commission. The deadline is January 15, 1986. Please send a letter of intent immediately to Ron Nieberding, Democratic Policy Commission, 499 South Capitol St., No. 422, Washington, D.C. 20003 (telephone 202+863–8162).

The second annual meeting of the Economists of New Jersey (ENJ) will be held March 22, 1986, at Trenton State College. Seminars and panel discussions will focus on The Economy of New Jersey; National Interest? Private vs. Public Services; and Encouraging the World Economy. For information and participation forms, contact Annette Meyer, President, ENJ, Department of Economics, Trenton State College, Trenton, NJ 08625.

Donation to nonprofit organization or institution: complete set of *American Economic Review*, December 1949 to present; *Journal of Economic Literature*, March 1969 to present. Contact Roslyn Arnold, 6800 Liberty Road, Apt. 514, Baltimore, MD 21207 (telephone 301+944–8239).

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants from the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for application to be received in the ACLS office are for meetings scheduled between July and October, March 1; meetings scheduled between November and February, July 1; meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Att: Travel Grant Program), 800 Third Avenue, New York, NY 10022, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

Deaths

Grant N. Farr, professor of economics, Pennsylvania State University, May 9, 1985.

Simon S. Kuznets, Cambridge, Massachusetts, July 8, 1985.

Retirements

John Due, professor of economics, University of Illinois-Urbana, August 20, 1985.

Leland C. Lehman, professor of economics, Bluffton College, May 31, 1985.

Joseph P. McKenna, professor of economics, University of Missouri-St. Louis, September 1, 1985.

Richard Ruggles, Stanley Resor professor of economics, Yale University, December 31, 1984.

Foreign Scholars

Mukul A. Asher, National University of Singapore: University Center for International Studies, University of Pittsburgh, August, 1985.

Koichi Hamada, University of Tokyo: visiting professor of economics, Yale University, September 1985.

Maria-Jose Herrero, London School of Economics: department of economics, University of Pittsburgh, September 1985.

Ron Holzman, CORE (Louvain): visiting assistant professor of economics, Virginia Polytechnic Institute and State University, 1985–86.

Bezalel Peleg, Hebrew University: visiting professor of economics, Virginia Polytechnic Institute and State University, spring 1986.

Marilda Sotomayor, Pontificio Universidade Catolica, Rio de Janeiro: department of economics, University of Pittsburgh, March 1985.

Promotions

Jean W. Adams: professor of economics, Iowa State University, July 1, 1985.

Lee J. Alston: associate professor of economics, Williams College, July 1985.

Richard A. Ashley: associate professor of economics, Virginia Polytechnic Institute and State University, fall 1985.

Steven Balkin: associate professor of economics, Roosevelt University, September 1, 1985.

Andrea H. Beller: associate professor, University of Illinois-Urbana, August 21, 1985.

Jan Brueckner: full professor of economics, University of Illinois-Urbana, August 21, 1985.

James H. Cassing: professor of economics, University of Pittsburgh, September 1, 1985.

Vincent P. Crawford: professor, department of economics, University of California-San Diego, July 1, 1985.

Jacques Crémer: professor, department of economics, Virginia Polytechnic Institute and State University, fall 1985.

Arthur T. Denzau: professor, department of economics, Washington University-St. Louis, July 1, 1985.

Sanjay Dhar: senior economist, external financing department, Federal Reserve Bank of New York, May 2, 1985.

Walter Enders: professor of economics, Iowa State University, July 1, 1985.

Lydia P. Harris: associate professor of economics, Goucher College, May 28, 1985.

Marvin L. Hayenga: professor of economics, Iowa State University, July 1, 1985.

Susan Hickok: senior economist, Special Studies Staff, external financing department, Federal Reserve Bank of New York, April 4, 1985.

R. Bryce Hool: professor of economics, SUNY-Stony Brook, September 1, 1984.

Robert W. Jolly: professor, department of economics, Iowa State University, July 1, 1985.

Tryphon Kollintzas: associate professor, department of economics, University of Pittsburgh, September 1, 1985.

Gary A. Latanich: associate professor of economics, Arkansas State University, July 1, 1985.

David L. Lindauer: associate professor, department of economics, Wellesley College, September 1985.

Robert McCauley: chief, Developing Economies Division, external financing department, Federal Reserve Bank of New York, May 2, 1985.

William H. Meyers: professor of economics, Iowa State University, July 1, 1985.

E. Dwight Phaup: professor of economics, Union College, September 1, 1984.

Allen Proctor: senior economist, Regional Economics Staff, monetary research department, Federal Reserve Bank of New York, April 4, 1985.

Salim Rashid: full professor, department of economics, University of Illinois-Urbana, August 21, 1985.

Eli Remolona: economist, Special Studies Staff, external financing department, Federal Reserve Bank of New York, April 22, 1985.

Mark J. Robert: associate professor of economics, Pennsylvania State University, July 1, 1985.

Adam Z. Rose: professor of mineral resource economics, West Virginia University, August 15, 1985.

Larry W. Samuelson: professor of economics, Pennsylvania State University, July 1, 1985.

David Shapiro: associate professor of economics, Pennsylvania State University, July 1, 1985.

Kenneth Stone: professor of economics, Iowa State University, July 1, 1985.

Loren W. Tauer: associate professor of agricultural economics, Cornell University, July 1, 1985.

T. Nicolaus Tideman: professor, department of economics, Virginia Polytechnic Institute and State University, fall 1985.

Mark Walker: professor of economics, SUNY-Stony Brook, September 1, 1984.

Administrative Appointments

Stanley W. Black: chairman, economics department, University of North Carolina-Chapel Hill, July 1, 1985.

Roger K. Chisholm, Fogelman College of Business and Economics, Memphis State University: dean, College of Business Administration, University of Arkansas-Little Rock, July 1, 1985.

Maximo V. Eng: chairman, department of economics and finance, St. John's University, July 1, 1985.

Ronald L. Friesen: chairman, department of economics and business, Bluffton College, June 1, 1985.

Jane V. Hall: chair, department of economics, California State University-Fullerton, fall 1985.

Andrew Hau: chairperson, department of economics, Millersville University, September 1, 1984.

Peter A. Prosper, Jr.: chair, department of economics, Union College, September 1, 1985.

Chong Soo Pyun: chairman, department of finance, insurance, and real estate, Memphis State University, July 1, 1985.

James D. Rodgers: head, department of economics, Pennsylvania State University, July 1, 1985.

Kevin C. Sontheimer: chair, economics department, University of Pittsburgh, September 1, 1985.

Thomas H. Tietenberg: chairman, department of economics, Colby College, July, 1985.

Thomas F. Torries: chairman, department of mineral resource economics, West Virginia University, July 1, 1985.

Appointments

Kabir Ahmad, University of Petroleum and Minerals, Saudia Arabia: associate professor of economics, Pennsylvania State University, August 19, 1985.

Rabah Amir, Yale University: assistant professor, joint appointment, department of applied mathematics and department of economics, SUNY-Stony Brook, September 1, 1985.

Laurence Ball, MIT: assistant professor of economics, Graduate School of Business Administration, New York University, September 1985.

Michel Balniski, Laboratoire d'Econometrie-Ecole Polytechnique, Paris: professor, joint appointment, department of applied mathematics and department of economics, SUNY-Stony Brook, September 1, 1985.

Taradas Bandyopadhyay, York University: associate professor of economics, Pennsylvania State University, August 19, 1985.

Jagdeeph Bhandari: associate professor of economics, West Virginia University, August 15, 1984.

David M. Blau, University of Miami: assistant professor of economics, University of North Carolina-Chapel Hill, July 1, 1985.

Jorge Brandts, University of Pennsylvania: assistant professor, department of economics, University of Connecticut, September 1, 1985.

Dale S. Bremmer: assistant professor of economics, Arkansas State University, August 15, 1985.

James N. Brown, Princeton University: assistant professor of economics, SUNY-Stony Brook, September 1, 1985.

Willem H. Buiter, London School of Economics: professor of economics, Yale University, July 1985.

Richard T. Carson, Jr.: assistant professor, department of economics, University of California-San Diego, July 1, 1985.

John J. Casson, American Express Company: assistant professor of economics and management science, Kean College of New Jersey, September 1, 1985.

Rick L. Chaney, Emory University: visiting assistant professor, economics department, University of North Carolina-Chapel Hill, July 1, 1985–June 30, 1986.

Christopher Cornwell: assistant professor of economics, West Virginia University, August 15, 1985.

Richard Cothren, Brown University: assistant professor, department of economics, Virginia Polytechnic Institute and State University, fall 1985.

Barbara Craig, University of Minnesota: assistant professor, department of economics, Virginia Polytechnic Institute and State University, fall 1985.

John B. Crihfield, University of Chicago: assistant professor of economics, Wake Forest University, fall 1985.

Adriaan Dierx: assistant professor of economics, West Virginia University, August 15, 1984.

Stewart Dorsey: associate professor of economics, West Virginia University, August 15, 1984.

Pradeep Dubey, Yale University: joint appointment, department of applied mathematics and department of economics, SUNY-Stony Brook, September 1, 1985.

Michael D. Duffy: assistant professor, department of economics, Iowa State University, July 1, 1985.

Michael Einhorn, Yale University: assistant professor, Rutgers University-Newark, September 1985.

John Fender, University of Lancaster, U.K.: associate professor of economics, Pennsylvania State University, August 19, 1985.

Richard Fowles, University of Utah: assistant professor, Rutgers University-Newark, September 1985.

Daniel Friedman: associate professor of economics, University of California-Santa Cruz, July 1, 1985.

James W. Friedman, Virginia Polytechnic Institute and State University: professor of economics, University of North Carolina-Chapel Hill, July 1, 1985.

Therese Garcia-Mila, University of Minnesota: assistant professor of economics, SUNY-Stony Brook, September 1, 1985.

John D. Graham, Carnegie-Mellon University: assistant professor of policy and decision sciences, Harvard School of Public Health, September 1, 1985.

Edward J. Green: professor, department of economics, University of Pittsburgh, September 1985.

Vassilis A. Hajivassiliou, MIT: assistant professor of economics, Yale University, July 1985.

David R. Hakes, Iowa State University: assistant professor of economics, University of Missouri-St. Louis, August 26, 1985.

Joseph Halevi, University of Montreal: visiting professor, department of economics, University of Connecticut, September 1, 1985.

Hans Haller, University of Mannheim: assistant professor, Virginia Polytechnic Institute and State University, fall 1985.

Gary D. Hansen: acting assistant professor, department of economics, University of California-Santa Barbara, July 1, 1985.

Panos Hatzipanayotou: assistant professor, department of economics, University of Connecticut, September 1, 1985.

Clifford B. Hawley, West Virginia University: visiting associate professor, department of economics and East-West Population Institute, University of Hawaii-Manoa, August 1, 1985.

Charles E. Hegji, Clinch Valley College: assistant professor, department of economics, Auburn University-Montgomery, June 1985.

Walter Hettich: professor, department of economics, California State University-Fullerton, fall 1985.

Catherine B. Hill, The Brookings Institution: assistant professor of economics, Williams College, July 1985.

Michael Hutchison: assistant professor of economics, University of California-Santa Cruz, July 1, 1985.

Jin-Ho Jeong, Indiana University: assistant professor of economics, University of Cincinnati, September 1, 1985.

Donald L. Johnson: assistant professor of economics, Arkansas State University, August 15, 1985.

Stanley R. Johnson: professor, department of economics, Iowa State University, March 15, 1985.

Boyan Jovanovic, New York University: visiting professor of economics, SUNY-Stony Brook, September 1, 1985.

William W. Lang, Yale University: assistant professor of economics, Rutgers University, September 1, 1985.

Dwight Lee, George Mason University: professor of economics, University of Georgia, September 1985.

Karen Lewis, University of Chicago: assistant professor of economics and international business, Graduate School of Business Administration, New York University, September 1985.

Duncan C. MacRae: Deputy Assistant Secretary for Economic Affairs, U.S. Department of Housing and Urban Development, January 20, 1985.

Michael Mandel, Harvard University: assistant professor of economics, Graduate School of Business Administration, New York University, September 1985.

Rajnish Mehra: associate professor, department of economics, University of California-Santa Barbara, July 1, 1985.

Douglas Mitchell: associate professor of economics, West Virginia University, August 15, 1984.

Leonard I. Nakamura, Princeton University: assistant professor of economics, Rutgers University, September 1985.

Abraham Neyman, Hebrew University: joint appointment, department of applied mathematics and department of economics, SUNY-Stony Brook, September 1, 1985.

Morteza Rahmatian: assistant professor of economics, West Virginia University, August 15, 1984.

James Rauch, Yale University: assistant professor of economics, SUNY-Stony Brook, September 1, 1985.

Thomas G. Rawski: professor, department of economics, University of Pittsburgh, September 1985.

David C. Rose, University of Virginia: assistant professor of economics, University of Missouri-St. Louis, August 26, 1985.

Peter Rupert: assistant professor of economics, West Virginia University, August 15, 1985.

Lewis Sage, Skidmore College: visiting professor of economics, Union College, September 1985.

Ryuzo Sato, Brown University: professor of economics, and director, The Center of Japanese Business and Economics, Graduate School of Business Administration, New York University, September 1985.

Selig Sechzer, University of Pennsylvania: assistant professor of economics, Rutgers University, September 1985.

Carlos Seiglie, University of Chicago: assistant professor, Rutgers University-Newark, September 1985.

W. Douglass W. Shaw, Jr., University of Colorado-Boulder: visiting professor of economics, Williams College, 1985–86.

Padmanabhan Srinagesh, University of Illinois-Chicago: assistant professor of economics, Williams College, July 1985.

Patricia L. Stuart, University of Colorado-Boulder: assistant professor of economics, Williams College, July 1985.

Martin A. Sullivan, Northwestern University: assistant professor of economics, Rutgers University, September 1985.

Eileen Trzcinski, University of Michigan: assistant professor, department of economics, University of Connecticut, September 1, 1985.

Shunichi Tsutsui, Brown University: assistant professor of economics, University of Georgia, September 1985.

John K. Wakeman-Linn, University of Wisconsin-Madison: assistant professor of economics, Williams College, July 1985.

Ronald Warren, University of Virginia: associate professor of economics, University of Georgia, September 1985.

Philip K. Way, Harvard University: assistant professor of economics, University of Cincinnati, September 1, 1985.

Paul Wilson, Brown University: assistant professor of economics, University of Georgia, September 1985.

Ann Dryden Witte, University of North Carolina: professor of economics, Wellesley College, September 1985.

John H. Wood, Northwestern University: professor of economics, Wake Forest University, fall 1985.

Makoto Yano, Cornell University: assistant professor of economics, Rutgers University, September 1985.

Leaves for Special Appointments

John S. Akin, University of North Carolina-Chapel Hill: World Bank, July 1, 1985–June 30, 1986.

Gerald D. Jaynes, Yale University: Study Director, Committee on Status of Black Americans, National Research Council, July 1985.

Stephen M. Miller, University of Connecticut: resident director, Program in European Studies, J.F. Kennedy Institute, Tilburg University, The Netherlands, September 1, 1985.

Resignations

Christophe Chamley, Yale University: The Hoover Institute, July 1985.

Russell W. Cooper, Yale University: Iowa State University, July 1985.

Zvi Eckstein, Yale University: Tel-Aviv University, July 1985.

Paul Ong, University of California-Santa Cruz: University of California-Los Angeles, July 1, 1985.

Jennifer Roback, Yale University: George Mason University, July 1985.

Marshall Robinson, president, Russell Sage Foundation, December 31, 1985.

Robert Rossana, The Pennsylvania State University, June 30, 1985.

Frederick R. Warren-Boulton, Washington University-St. Louis, July 1, 1985.

Robert J. Willis, SUNY-Stony Brook: University of Chicago and National Opinion Research Center, August 31, 1985.

Brian D. Wright, Yale University: University of California-Berkeley, July 1985.

Miscellaneous

Francis X. Sutton: acting president, Social Science Research Council, October, 1985.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, use the following style:

A. Please use the following categories (please—do not send public relation releases):

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), new institution and the date at which the change will occur.

C. Type each item on a separate 3×5 card.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

All items and information should be sent to the Assistant Editor, *American Economic Review*, 169 Nassau Street, Princeton, NJ 08542–7067.

NOTICE TO ALL GRADUATE DEPARTMENTS

The December 1986 issue of the *Review* will carry the eighty-third list of doctoral dissertations in political economy in American universities and colleges. The list will give recipients and titles of doctoral degrees conferred during the academic year terminating June 1986. This announcement is an invitation to send us information for the preparation of the list.

By June 30, please send us this information on 3×5 cards, conforming to the style shown below, one card for each individual. Please indicate by a classification number in the right-hand corner the field in which the thesis should be classified. The classification system is that used by the *Journal of Economic Literature* and printed in every issue.

All items and information should be sent to the Editorial Office, *American Economic Review*, 169 Nassau Street, Princeton, NJ 08542–7067.

JEL Classification No. ________

Name: LAST NAME IN CAPS: First Name, Initial

Institution Granting Degree: ________________________

Degree Conferred (Ph.D. or D.B.A.) ____________ Year ____________

Dissertation Title: ________________________

EIGHTY-SECOND LIST OF DOCTORAL DISSERTATIONS IN POLITICAL ECONOMY IN AMERICAN UNIVERSITIES AND COLLEGES

The present list specifies doctoral degrees conferred during the year terminating June 1985.

General Economics; including Economic Theory, History of Thought, Methodology, Economic History, and Economic Systems

JOSE M. AIZPURUA, Ph.D. Minesota 1985. Coordination equilibria among oligopolists.

NAHID ASLANBEIGUI, Ph.D. Michigan 1984. Modern welfare economics and positive science.

JOHN C. ATTARIAN, Ph.D. Michigan 1984. Failure of a vision: Critique of objectivist economic philosophy.

RONALD J. BALVERS, Ph.D. Pittsburgh 1984. Uncertainty, aggregate supply, and disequilibrium.

ANINDA KUMAR BOSE, Ph.D. SUNY-Stony Brook 1985. A new look at the economics of quality.

STUART S. BROWN, Ph.D. Columbia 1985. "Artificial pricing" and marketing disruption by nonmarket economies.

JONATHAN BURKE, Ph.D. MIT 1985. Essays on equilibria in dynamic economies.

KELLY D. BUSCHE, Ph.D. Simon Fraser 1985. Toward a theory of repeating Dutch auctions.

JOHN Y. CAMPBELL, Ph.D. Yale 1984. Asset duration and time-varying risk premia.

YANG MING CHANG, Ph.D. SUNY-Buffalo 1985. Three essays of optimal enforcement and protection.

CHIEN-FU CHOU, Ph.D. Yale 1984. Two essays on game theory.

ROBERT CLAIR, Ph.D. Pennsylvania 1985. A model of bank behavior under uncertainty: An expansion and test of the Mason model.

JEFFREY L. COLVIN, Ph.D. American 1984. Fixed capital and the labor theory of value.

KEVIN D. COTTER, Ph.D. Minnesota 1984. Agents' information as a market variable: New techniques.

ANNA P. DELLA VALLE, Ph.D. Columbia 1984. Intertemporal pricing for the use of capital in production and its relationship to investment and replacement policy.

TIMOTHY M. DEVINNEY, Ph.D. Chicago 1984. A general equilibrium analysis of the borrower-lender relation: An examination of credit rationing hypotheses.

NESTOR DOMINGUES, Ph.D. Maryland 1984. Two essays on the Q model of investment.

ELMA A. M. DROSSO, Ph.D. Simon Fraser 1985. The fall and rise of labor dues: The thirteenth-century manor in the shadow of the king.

MARC P. DUDEY, Ph.D. Princeton 1984. Essays on monopoly pricing strategies.

DAVID W. EMMONS, Ph.D. Johns Hopkins 1985. Essays on the existence of equilibria in measure theoretic public goods economies.

MAXIM P. ENGERS, Ph.D. California-Los Angeles 1984. The existence and uniqueness of signalling equilibria.

ROY J. EPSTEIN, Ph.D. Yale 1984. Econometric methodology in historic perspective.

JERRY EVENSKY, Ph.D. Syracuse 1985. Expanding the neoclassical vision—Recapturing the lost legacy of Adam Smith.

ELIZABETH B. FIELD, Ph.D. Duke 1985. Elasticities of complementarity and returns to scale in antebellum cotton agriculture.

BEVERLY FLEISHER, Ph.D. Michigan State 1985. A new measure of attitude towards risk.

HIDEKI FUNATSU, Ph.D. Southern Methodist 1984. Theory of export credit insurance.

RAUL E. GARCIA, Ph.D. Washington 1985. Inflationary expectations and the term structure of interest rates: Argentina, 1977–82.

JASPAL S. GHATHA, Ph.D. New School 1985. On the significance of the assumption of returns in economic analysis.

K. GARY GRANT Ph.D. Western Ontario 1984. The Canadian wheat trade during the interwar years.

HENRI GROENEVELT, Ph.D. Columbia 1985. Resource allocation problems with decreasing marginal returns to scale.

FARLEY GRUBB, Ph.D. Chicago 1984. Immigration and servitude in the colony and Commonwealth of Pennsylvania: A quantitative and economic analysis.

DAVID L. HAMMES, Ph.D. Simon Fraser 1985. Prices, information, and explanation in neoclassical economic theory.

ROBERT G. HANSEN, Ph.D. California-Los Angeles 1984. Theory and evidence on the consequences of alternative auction rules.

DONALD B. HAUSCH Ph.D. Northwestern 1985. A common value auction model allowing asymmetrically informed bidders, risk-adverse bidders, and a comparison of sequential versus simultaneous auctions for multiple objects.

BERNARDUS J. HEIJDRA, Ph.D. Simon Fraser 1985. Empirical and theoretical aspects of non-Walrasian economics.

YOLANDA HENDERSON, Ph.D. Pennsylvania 1984. A general equilibrium taxation model with asset choice.

DAVID HIRSHLEIFER, Ph.D. Chicago 1985. Risk, equilibrium, and futures markets.

SEONG H. HONG, Ph.D. Columbia 1985. Essays in transfer, taxation, and the limit of income redistribution policy in market economies.

CHEN-MIN HSU, Ph.D. Johns Hopkins 1985. A study in the theory of temporary equilibrium with quantity rationing.

DOO HYUN HWANG, Ph.D. Ohio State 1984. Hedonic estimation of the demand for housing: A theoretically based selection method for the functional form.

FREDERIC B. JENNINGS, JR., Ph.D. Stanford 1985. Public policy, planning horizons, and organizational breakdown: A postmortem on British canals and their failure.

JASON S. JOHNSTON, Ph.D. Michigan 1984. Imperfect information and the legal process: Toward a general economic theory of tort liability.

SUNGHEE JWA, Ph.D. California-Los Angeles 1983. Towards an equilibrium approach to the effects of price controls: A theoretical and empirical analysis of the effects of price controls on quality offerings.

BARBARA KAHN, Ph.D. Columbia 1984. Measuring variety-seeking and reinforcement behavior using panel data.

NANCY F. KANE, Ph.D. California-Los Angeles 1984. Industry relocation and wage equalization: The case of U.S. textiles, 1880–1930.

NICHOLAS V. KARAMOUZIS, Ph.D. Pennsylvania State 1984. The role of "openness" in the conduct of U.S. monetary policy.

ORHAN KAYAALP, Ph.D. City-New York 1984. Ugo Mazzola and the Italian doctrine of finance.

RAHUL KHULLAR, Ph.D. Boston 1985. Adverse selection, certification, and quality standards.

MOO-HONG KIM, Ph.D. Southern California 1984. Studies of behavioral axioms for comparative status under uncertainty.

CARL A. KOGUT, Ph.D. Texas A&M 1984. Individual and market bidding behavior in Vickrey first-price auctions.

SUNNY K. S. KWONG, Ph.D. British Columbia 1985. Price-sensitive inequality measure.

URHO V. K. LEMPINEN, Ph.D. Princeton 1984. Optimizing agents, exogenous shocks, and adjustments in the economy: Real and nominal fluctuations in economies with a wage rigidity.

DAVID LENZE, Ph.D. Johns Hopkins 1985. Stochastic factor supply and the theory of the firm.

LARRY S. LEV, Ph.D. Michigan State 1984. The role of flexibility in the design, acquisition, and use of durable inputs.

MARK LEVINSON, Ph.D. New School 1985. The rise and decline of classical political economy.

JOHN K. LODEWIJKS, Ph.D. Duke 1985. Arthur Okun and American macroeconomics.

GEORGE F. LOEWENSTEIN, Ph.D. Yale 1985. Expectations and intertemporal choice.

KEVIN MCCABE, Ph.D. Pennsylvania 1984. Essays on random meeting economies.

DONALD N. MACDONALD, Ph.D. Texas A&M 1984. Experimental studies of individual choice behavior in uncertain contexts.

BRIAN H. MCGAVIN, Ph.D. California-Los Angeles 1985. An innovation analysis of structural change in consumption.

W. F. MCKENNA, Ph.D. Temple 1985. An empirical study of the comparative external validity of alternative approaches to the measurement of consumer utility functions.

DORIS MAHAFFEY, Ph.D. North Carolina-Chapel Hill 1985. The Yugoslav economy: A case study of a labor-managed economic system.

NICHOLAS G. MANKIW, Ph.D. MIT 1984. Essays on consumption.

DUCAN MANN, Ph.D. Pennsylvania 1984. Essays in microeconomic theory.

NORIATSU MATSUI, Ph.D. Ohio State 1984. Economic justice and income distribution: Towards a theory of distributive justice.

PHILIP T. MINTZ, Ph.D. Maryland 1985. Allocation of gains by group behavior among group members: New theory of bargaining.

JEFFREY A. MIRON, Ph.D. MIT 1984. The economics of seasonal time-series.

ARNOLD G. MOSKOWITZ, Ph.D. New York 1985. Examination of inflation effects on interest rates.

EDWARD NOSAL, Ph.D. Queen's 1985. Incomplete markets and implicit contract theory.

DARRELL F. PARKER, Ph.D. Purdue 1984. Price and output dynamics: Microfoundations, timing, and transmission.

IVAN P. L. P'NG, Ph.D. Stanford 1985. Information and the law.

BRIAN T. PRATT, Ph.D. Simon Fraser 1985. Two papers: *Ex post* exchange activity and the existence of risk affinity, and learning by doing and the structure of employment contracts under uncertainty.

ERNEST RAIKLIN, Ph.D. New School 1985. Characteristics and peculiarities of contemporary Soviet commodity production.

WALTER T. RHODUS, Ph.D. Ohio State 1985. Allocative efficiency of experimental markets under conditions of supply and demand uncertainty.

CHRISTINA D. ROMER, Ph.D. MIT 1985. The instability of the prewar economy reconsidered: A critical examination of historical macroeconomic data.

P. NICHOLAS ROWE, Ph.D. Western Ontario 1985. Rules and institutions.

LEWIS C. SAGE, Ph.D. Maryland 1985. The role of information in strategic voting decisions.

TIM R. SASS, Ph.D. Washington 1984. The economics of information intermediaries.

PAUL S. SEGERSTROM, Ph.D. Rochester 1985. The process of creative destruction and stochastic games.

BARRY J. SELDON, Ph.D. Duke 1985. A nonresidual approach to the measurement of social returns to research with application to the softwood plywood industry.

MARILYN H. SIMONS, Ph.D. SUNY-Stony Brook 1984. Bidding with a budget constraint.

KENNETH A. SNOWDEN, Ph.D. Wisconsin-Madison 1984. Three essays on the American capital market, 1870–1913.

WILLIAM E. SPRIGGS, Ph.D. Wisconsin-Madison 1984. Afro-American wealth accumulation; Virginia, 1900–14.

BALAKRISHNAN SUBRAMANIAN, Ph.D. Columbia 1985. Evolutionary aspects of monetary theories of the business cycle.

LEE F. SUNDHOLM, Ph.D. Northern Illinois 1985. Protectionism in a period of mercantilism: The English experience from the reign of Elizabeth I to Adam Smith's *Wealth of Nations*.

GUIDO E. TABELLINI, Ph.D. California-Los Angeles 1984. The formation of expectations and central bank behavior.

TOMMY CHIN-CHIU TAN, Ph.D. Princeton 1984. Information, repeated interaction, and credibility.

ELENI THANOU, Ph.D. Clark 1985. Symmetry in production functions: Implications for theory and empirical analysis.

LYNDA THOMAN, Ph.D. Stanford 1984. Purchases under quality uncertainty.

ARJA H. TURUNEN, Ph.D. British Columbia 1985. Optimal policy choices in open economies.

FERNANDO VEGA-REDONDO, Ph.D. Minnesota 1984. On the optimal allocations of public goods and externalities in nonconvex economies: Statics, dynamics, and incentives.

MARION D. WATTS, Ph.D. Alabama 1985. Corporatism and Roman Catholic socioeconomic thought: An analysis of current issues.

ROBERT B. WILLIAMS, Ph.D. North Carolina-Chapel Hill 1984. Inside the labor-managed firm: An economic analysis.

MING-JER WU, Ph.D. Connecticut 1985. The optimal fiscal and monetary policy mix in a rational expectations model.

KENJI YAMAMOTO, Ph.D. SUNY-Buffalo 1984. Pareto-efficient allocations and competitive equilibrium in overlapping-generations models.

STEPHEN P. ZELDES, Ph.D. MIT 1984. Optimal consumption with stochastic income.

Economic Growth and Development; including Economic Planning Theory and Policy; Economic Fluctuation and Forecasting

MOHAMED ABDULLA, Ph.D. Boston 1985. An econometric model of a low-income economy with particular reference to Kenya: Analysis of the behavioral relations among the major macroeconomic aggregates.

SOLOMON IZIELEN AGBON, Ph.D. Texas-Austin 1985. Class and economic development in Nigeria, 1900–80.

TAHMOURES AHMADI AFSHAR, Ph.D. Indiana-Bloomington 1984. Private returns and social costs and benefits of rural-urban migration in less developed countries: A case study of India, 1961-71.

SAIYED F. AL-KHOULI, Ph.D. Colorado-Boulder 1985. Forecasting of the active skills by occupation and nationality in a petrodollar surplus economy: The case of Saudi Arabia.

ALOYS B. AYAKO, Ph.D. Boston 1985. Kenya farm household consumption and production behavior: A policy simulations model.

OSAMA BAHANSHAL, Ph.D. Colorado State 1984. Saudi Arabia's domestic crude oil production requirements: A physical and monetary approach.

RIFAAT K. BASANTI, Ph.D. Northwestern 1985. The monetary approach to structural adjustment problems in less developed countries: A case study of Sudan.

PETER A. BERMAN, Ph.D. Cornell 1984. Equity and cost in the organization of primary health care in Java, Indonesia.

DEEPAK BHATTASALI, Ph.D. Boston 1985. External debt and balance-of-payments adjustment in some non-oil-exporting less developed countries.

ABUDLLAH BIN MOHD TAHIR, Ph.D. American 1984. Tax and investment incentives in the developing countries: A case study of the manufacturing sector of Malaysia.

MELISSA H. BIRCH, Ph.D. Illinois-Urbana 1985. Public enterprise in electric power provision: Its role in Paraguayan economic development.

PRASARN BOONSERM, Ph.D. Notre Dame 1985. The potential impact of Thailand's alcohol program on production, consumption, and trade of cassava, sugarcane, and corn.

PAUL BURKETT, Ph.D. Syracuse 1984. Savings mobilization in the Third World: Theory and evidence from Peru.

CHANGHEE CHAE, Ph.D. Columbia 1984. LDC's external borrowing and debt servicing capacity.

RICK L. CHANEY, Ph.D. Illinois-Urbana 1984. The economics of 100 years of emigration and remittances in Portugal.

ELMER P. CHASE III., Ph.D. International 1984. A long wave explanation of the inflation of the late 1960's and 1970's in the United States.

YUN-HSING CHEUNG, Ph.D. Oregon 1984. Technology transfer, resources import, and economic growth of newly industrializing countries.

RONYOOT CHITRADON, Ph.D. SUNY-Buffalo 1985. An allocation of resources of hospital care in developing countries.

YEE YEN CHUA, Ph.D. California-Santa Barbara 1984. Wage differentials in Peninsular Malaysia.

FELIX M. CUE, Ph.D. Oklahoma State 1985. Income distribution and economic development: A case

study of the Kuznets hypothesis applied to Puerto Rico.

ELIEZER CURET, Ph.D. American 1984. Development policies for small countries: The case of Puerto Rico.

EDNALDO DA SILVA, Ph.D. California-Berkeley 1985. Peasant production, labor reserve, and the food economy of northeast Brazil.

NURHAN DAVUTYAN, Ph.D. California-Santa Barbara 1983. Price stabilization policies: Robustness and asymmetry.

JEAN-JACQUES DETHIER, Ph.D. California-Berkeley 1985. The political economy of food prices in Egypt.

WILLIAM R. EASTERLY, Ph.D. MIT 1985. A computable general equilibrium model of Mexico with portfolio balances: With application to devaluation.

CHINYERE E. EGBE, Ph.D. Washington State 1984. The impact of foreign private investment on the growth of *GNP* and investment in Nigeria.

HASSABEL R. A. ELBASHIR, Ph.D. Indiana-Bloomington 1984. A monetary approach to the balance of payments under fixed exchange rate: The case of Sudan.

HADI S. ESFAHANI, Ph.D. California-Berkeley 1984. A system-wide analysis of the impact of policy on agricultural performance in Egypt.

MARK L. GARDNER, Ph.D. Georgia State 1985. Federal government deficits and the interaction of fiscal and monetary policies: An aggregate demand/aggregate supply model simulation.

GUSTAVO P. GOMES, Ph.D. Illinois-Urbana 1985. State intervention in the contemporary Brazilian economy.

DANIEL L. GRESSEL, Ph.D. Chicago 1984. The effects of trade policy, size of government, and inflation on growth rates in less developed countries.

ABDULWAHED K. HAMED, Ph.D. Wisconsin-Milwaukee 1984. Capital absorptive capacity of nonoil sectors in Saudi Arabia.

ELMORZY E. HEGAZEY, Ph.D. Connecticut 1985. The contribution of oil to the economic development of Kuwait (1962–81): A macroeconometric approach.

JERRY HEMBD, Ph.D. Stanford 1984. The political economy of uneven regional development in the Philippines: The case of Bikol.

MAURICE R. HOPPIE, Ph.D. Tennessee-Knoxville 1985. The Mahaica-Mahaicony-Abary Project: Socioeconomic change in Guyana.

ESMAIL HOSSEINZADEH, Ph.D. New School 1985. An evaluation of the theory of noncapitalist development: The Soviet program for the Third World transition to socialism.

BEEN-KWEI HWANG, Ph.D. Oklahoma 1984. A dynamic simultaneous model of Taiwan: IS-LM, AS-AD, dynamic multiplier analysis, business cycle, forecast, simulation, and policy experiment.

KOH HYUNWOOK, Ph.D. Connecticut 1985. A quantitative analysis of entrepreneurship: The Korean case.

BARRY W. ICKES, Ph.D. California-Berkeley 1984. A macroeconomic theory of centrally planned economics.

RANDALL S. JONES, Ph.D. Michigan 1984. The economic development of colonial Korea.

FADWA KALAJI, Ph.D. Pittsburgh 1985. Analysis of savings behavior in the countries of the Arab world.

CHUL K. KANG, Ph.D. Northwestern 1985. Effects of a monopolistic market on the growth pattern and the income distribution of a growing two-sector economy.

SAMAILA KEWA, Ph.D. SUNY-Binghamton 1984. Alternative food security policies for Nigeria: A dynamic-stochastic simulation approach.

CHONG KU KIM, Ph.D. California-Berkeley 1984. Adjustment rules and planning for development strategies under international market price uncertainty.

VLADIMIR KONTOROVICH, Ph.D. Pennsylvania 1985. Soviet productivity and technical change.

ROBERT B. KOOPMAN, Ph.D. Boston College 1984. Soviet project choice criteria: An empirical evaluation of the 1969 reforms.

PRADEEP KOTAMARAJU, Ph.D. Illinois-Urbana 1985. The role of noncommercial fuels in India's economic development.

TIMUR KURAN, Ph.D. Stanford 1982. Price adjustment under inflation and rules of trade: Two essays.

STEPHEN S. KYEREME, Ph.D. Cornell 1984. Food consumption and poverty patterns in Ghana.

MAN WOO LEE, Ph.D. Minnesota 1984. Optimal taxation on foreign investment in growing economies.

SANG MAN LEE, Ph.D. Claremont 1985. The controllability of the money supply in a developing economy: The case of Korea.

GARY J. MCMAHON, Ph.D. Western Ontario 1984. International commodity agreements and a small developing country: The case of Kenya and the international coffee agreement.

RAMON MARIMON, Ph.D. Northwestern 1984. General equilibrium and growth under uncertainty: The turnpike property.

SUGATA MARJIT, Ph.D. Rochester 1985. Three essays on trade of developing nations.

DONALD C. MASTERS, Ph.D. George Washington 1985. Export sector policies and performance: Guatemala coffee sector, 1961–81.

ADOLFO MEISEL, Ph.D. Illinois-Urbana 1984. Exchange rate policy in Colombia—1970's.

MERLE R. MENEGAY, Ph.D. Michigan State 1985. Improving the performance of procurement systems for fruit and vegetable processors in Thailand: A case study of up-country picklers and canneries.

ISAAC J. MINDE, Ph.D. Michigan State 1985. Economic analysis of farm and nonfarm rural employment in Morogoro District, Tanzania.

MASOUD MOGHADDAM, Ph.D. Iowa State 1984. Two essays on money supply growth, inflation, and interest rates.

YOUSUF H. MOHAMMAD, Ph.D. California-Santa Barbara 1983. Recent developments within OPEC: Economic perspective.

KAREN M. MOKATE, Ph.D. Illinois-Urbana 1984. The Colombian manufacturing industry during the era of OPEC price shocks.

ADIL H. MOUHAMMED, Ph.D. Nebraska-Lincoln 1985. Econometric model of development strategies in Iraq: 1951–80.

PANAYIOTIS MOURDOUKOUTAS, Ph.D. SUNY-Stony Brook 1985. Seasonal employment and tourist development: The case of the Greek islands.

FRANCIS M. MWEGA, Ph.D. Illinois-Urbana 1985. A general equilibrium analysis of tax incidence in a developing country: The case of Kenya.

MOHAMMED S. T. NAFIL, Ph.D. Washington State 1984. Investment performance, capital absorptive capacity, oil revenues, and economic development in an oil-based economy: The case of Iraq, 1951–78.

BHANUPONGSE NIDHIPRABHA, Ph.D. Johns Hopkins 1985. External shocks and adjustment policies in Thailand.

CHRISTOPHER W. NORDLINGER, Ph.D. Fletcher School 1984. The economic benefits of telecommunication in a developing country: Evaluating a rural satellite system for Senegal.

YONG-SUK OH, Ph.D. Indiana-Bloomington 1985. Modification of the Feldman growth model and its theoretical and policy reflection in the Chinese economy, 1953–80.

LINDA A. PAUL, Ph.D. Maryland 1985. The urban informal sector in economic development.

ALAN L. PIAZZA, Ph.D. Stanford 1985. Food consumption and nutritional status in the People's Republic of China.

GREGORY G. PICKETT, Ph.D. California-Santa Barbara 1983. An option valuation model of bonus bidding and profit-share bidding for offshore oil and gas leases.

ANTHONY P. RODRIQUES, Ph.D. California-Berkeley 1984. Short-run potential output growth with putty and clay.

FARHAD SABOORI, Ph.D. Indiana-Bloomington 1985. The short-run relationship between inflation and output changes in developing countries: The case of Latin America.

HOSSEIN SARJEHPEYMA, Ph.D. Oklahoma 1984. The economic impact of internal migration: Iran as special case.

SAVVAS SAVVIDES, Ph.D. New York 1985. The monetary sector and monetary policy in Cyprus: A monetary econometric model.

MOSCHOS M. A. SCOULLIS, Ph.D. New York 1985. The balance of payments of Greece: Its structural and macroeconomic aspects.

MOHAMED SHAMS, Ph.D. Texas-Austin 1984. Oil conservation and economic development in Saudi Arabia.

JEFFREY A. SHAPIRO, Ph.D. Purdue 1985. Forecasting aggregate economic activity: The usage of short-term and long-term leading indicators.

SANGDAL SHIM, Ph.D. Minnesota 1984. Inflation and the government budget constraint: International evidence.

BENCHAPHUN SHINAWATRA, Ph.D. Michigan State 1985. Highland-lowland interrelationships in northern Thailand: A study of production, distribution, and consumption.

LEE-YING SOON, Ph.D. Pennsylvania 1984. Some aspects of household and labor market in a developing country with emphasis on self-employment.

CHURAI TAPVONG, Ph.D. West Virginia 1985. Trade policy as an input to development: An analysis of export promotion policy in Thailand, 1972–80.

WILLIAM I. TIERNEY, Ph.D. Michigan State 1984. The structure, conduct and performance of the CCC export credit sales program with a case study of the additionality effect of CCC financed cotton exports to Korea.

GUSTAVO E. TREVINO, Ph.D. Texas-Austin 1985. External shocks and alternative economic policies under the "oil economy syndrome": A multisectoral model of Mexico using the transactions value approach.

VICTOR UKPOLO, Ph.D. American 1985. An economic evaluation of the Yobe River irrigation project in Nigeria.

EBENGE E. USIP, Ph.D. Connecticut 1985. A macroeconometric model of the Nigerian economy: A simulated analysis of oil shocks in a development context.

JORGE ALBERTO VASCONCELLOS, Ph.D. Columbia 1984. Strategic-key success factors and strengths.

TO-FAR WANG, Ph.D. California-Santa Barbara 1983. Econometric and input-output analysis of income distribution, growth, and employment: A case study of Taiwan.

WAYNE YIH-YUH WAUNG, Ph.D. Minnesota 1984. Inflation and output fluctuations in Taiwan.

JOHN R. WELLS, D.B.A. Harvard 1985. In search of synergy.

DAVID M. WISHART, Ph.D. Illinois-Urbana 1985. The political economy of conflict over water rights between Israel and Jordan.

DAVID C. WONG, Ph.D. California-Santa Barbara 1984. Development planning in a socialist labor-managed economy.

LUNG-FAI WONG, Ph.D. Minnesota 1985. A comparative analysis of agricultural productivity growth among socialist countries.

BEHZAD YAGHMAIAN, Ph.D. Fordham 1985. economic development land reform and imports substitution: The case of Iran.

Economic Statistics; including Econometric Methods, Economic and Social Accounting

RABAH AMIR, Ph.D. Illinois-Urbana 1985. Essays on nonclassical growth and differential games.

LAWRENCE M. AUSUBEL, Ph.D. Stanford 1984. Partially revealing equilibria.

MOHAMMAD S. BAJWA, Ph.D. Northern Illinois 1985. Determinants of U.S. inflation—ML and MINQU estimation of arch models.

SCOTT W. BARNHART, Ph.D. Texas A&M 1984. An investigation into the causes of nonmartingale behavior in commodity futures prices.

CHARLES E. BATES, Ph.D. Rochester 1985. Nonlinear parametric models with dependent observations.

JOHN T. BATTS, Ph.D. Colorado-Boulder 1985. Relative predictive ability of three autoregressive moving average models: A Monte Carlo investigation.

ARY L. BOVENBERG, Ph.D. California-Berkeley 1984. Capital accumulation and capital immobility: *Q*-theory in a dynamic general equilibrium framework.

GALE A. BOYD, Ph.D. Southern Illinois-Carbondale 1985. Scale and productivity in surface coal mining.

MERAL D. BUYÜKKURT, Ph.D. Indiana-Bloomington. Analysis of a two-echelon repairable item system with state dependent failure and repair rates.

CARLO CARRARO, Ph.D. Princeton 1985. New methods for macroeconomic policy analysis.

YU-MIN CHEN, Ph.D. Purdue 1984. Estimation of random coefficient dynamic models: Two applications.

CHIN YI CHU, Ph.D. Michigan 1985. Two essays on stochastic income theory.

ROBERT T. CLEMEN, Ph.D. Indiana-Bloomington 1984. Modeling dependent information: A Bayesian approach.

J. EDEN CLOUTIER, Ph.D. Queen's 1984. The estimation of time-varying parameter models with a Markov adjustment process.

MASAYOSHI FUSE, Ph.D. California-Berkeley 1985. An aggregate planning model for make-to-order production.

DAVID M. GARMAN, Ph.D. Michigan 1984. The theory of cross-equation shrinkage estimation and an application to Michigan SMSAs.

ERIC GHYSELS, Ph.D. Northwestern 1984. The economics of seasonality: The case of the money supply.

DANIEL V. GORDON, Ph.D. British Columbia 1984. Testing for structure in a multiproduct industry with price expectations: The Canadian cattle industry.

DAVID GRIFFITHS, Ph.D. Simon Fraser 1985. Some aspects of inflation and the variability of relative prices.

EUSTACE A. GUNTHROPE, Ph.D. Columbia 1985. Constrained estimation in nonlinear systems.

VASSILIS A. HAJIVASSILIOU, Ph.D. MIT 1985. Disequilibrium econometric modelling and related limited dependent variables models.

AARON K. HAN, Ph.D. Stanford 1984. Two essays in estimation methods using ordinal data.

JOSEPH G. HIRSCHBERG, Ph.D. Southern California 1984. The impact of time-of-use electricity pricing: A firm-level econometric approach.

JIRO HODOSHIMA, Ph.D. California-Berkeley 1984. Inference of limited dependent variables models: Parametric and nonparametric approaches.

SUSAN H. G. HOLLAND, Ph.D. Purdue 1984. An alternative measurement structure: The positively ordered partial semigroup with identity.

RANDY HOLLIDAY, Ph.D. West Virginia 1985. A model of disequilibrium adjustments between spot and futures markets.

MOHAMMAD-ALI KAFAEI, Ph.D. Michigan State 1984. On the adequacy of the Sargan approximation to the normal in econometric models.

ARTHUR KENNICKELL, Ph.D. Pennsylvania 1984. An investigation of life cycle savings behavior in the United States.

SHADY S. KHOLDY, Ph.D. California-Santa Barbara 1984. An efficiency test for Korean agricultural sector: A multi-output translog cost function, 1961–79.

KEUCK SOO KIM, Ph.D. Pittsburgh 1984. Nonsurvey input-output updating techniques: A comparison of accuracy with Korean data.

BEN L. KYER, Ph.D. West Virginia 1984. The role of real money balances in the U.S. manufacturing sector productivity slowdown, 1955–81.

YUN-BAI LEE, Ph.D. Indiana-Bloomington 1985. An algorithm for global minimization of a concave function with linear constraints.

PANG TIEN LIEU, Ph.D. Texas A&M 1985. A comparative study of the empirical impact on the joint estimation approach and the alternative selection approach for application to econometric models with a pure autocorrelation structure: The case of the Fechner-Thurstone direct utility function.

DENNIS R. MCLAIN, Ph.D. California-Berkeley 1984. A systems approach to the aeromedical aircraft routing problem using a computer-based model.

LONNIE J. MAGEE, Ph.D. Western Ontario 1984. Analytic studies in econometric inference with small samples.

CATHERINE A. MELFI, Ph.D. North Carolina-Chapel Hill 1984. Estimation and decomposition of productive efficiency in a panel data model: An application to electric utilities.

HANID MUKHTAR, Ph.D. Boston 1985. Essays in consumption behavior: Evidences Pakistani data.

THOMAS S. PAYNTER, Ph.D. California-Berkeley 1985. Neural mechanisms and economic models.

ROBERT F. PHILLIPS, Ph.D. Columbia 1985. The evaluation and combination of forecasts and a linear

programming approach to an empirical Bayes estimate of the conditional mean of a normal population.

ANDRE R. PLOURDE, Ph.D. British Columbia 1985. Modelling the economic implications of offshore oil: The case of Hibernia.

RAFIQUER RASUL, Ph.D. Cincinnati 1985. An approach to gold price behavior with reference to price expectations.

SRINIVAS REDDY, Ph.D. Columbia 1985. Effects of misspecification and strength of signal on the estimation in structural equation models.

RICHARD G. ROBB, JR., Ph.D. Chicago 1985. Three essays on the identification of economic models.

ALAN J. ROGERS, Ph.D. Princeton 1984. Essays on aspects of the theory of hypothesis testing in large samples.

PHADEJ ROJANASAKUL, Ph.D. Wisconsin-Milwaukee 1984. A study of the true cost-of-living index using alternative utility functions.

KATHLEEN E. S. SHORT, Ph.D. Michigan 1984. The life cycle theory, uncertainty, and the saving behavior of the elderly.

KESHAB SHRESTHA, Ph.D. SUNY-Buffalo 1984. Multiple indicator approach to the analysis of distributed lag of multiple causes.

RICHARD SIMES, Ph.D. Pennsylvania 1984. Issues in the econometrics of policy evaluation.

TARACAD R. SIVASANKARAN, Ph.D. New York 1985. Intelligent model management for an actuarial consulting system.

JAMES N. STANARD, Ph.D. New York 1985. A simulation test of prediction errors of loss reserve estimation techniques.

DAVID D. STEINER, Ph.D. Michigan 1984. Bayesian learning and tests of the rational expectations hypothesis.

ATHANSIOS STENGOS, Ph.D. Queen's 1984. Nonlinear estimation in models with limited dependent variables and sample selectivity.

BRUCE TABAKMAN, Ph.D. City-New York 1985. A frequency decomposition approach to exchange rate forecasting.

MITSUO TAKASE, Ph.D. SUNY-Buffalo 1985. Effect of unanticipated inflation on saving rates.

LARRY W. TAYLOR, Ph.D. North Carolina-Chapel Hill 1984. Testing for misspecification in the general linear and probit models.

EMSLEY TROMPE, Ph.D. Pennsylvania State 1985. The impact of regulation on electric utility pricing: An econometric test of Ramsey pricing.

MALCOLM E. TURNER, Ph.D. Simon Fraser 1984. Stochastic optimal control using a Canadian econometric model.

JONE-LIN WANG, Ph.D. Southern California 1984. The computation and the sensitivity analysis of economics equilibrium.

KYHYANG YUHN, Ph.D. Northern Illinois 1985. Intertemporal profit maximization and dynamics of investment: Developing the Hamiltonian representation of the normalized translog profit function.

Monetary and Fiscal Theory, Policy and Institutions

HARJIT K. ARORA, Ph.D. Wisconsin-Milwaukee 1984. Budget deficits and the real interest rate: A loanable funds approach.

CHIICHII ASHWE, Ph.D. Queen's 1985. Revenue sharing in Nigeria: A study of fiscal adjustment in a federal country.

NICHOLAS BARCIA, Ph.D. City-New York 1985. Short-term interest rates and inflation.

FRANK G. BARRY, Ph.D. Queen's 1985. Government policy in a small open economy with capital accumulation and classical unemployment.

L. J. BEGLEY, Ph.D. Temple 1985. U.S. commercial banks' tax position and their impact on South American lending.

ROBERT A. BLACK, Ph.D. West Virginia 1984. Economic slack and variable-weight reduced-form money multipliers for price and output.

MICHAEL T. BOND, Ph.D. Case Western Reserve 1985. Nominal interest rates, expected inflation, marginal tax rates, and inflation uncertainty.

ELIJAH BREWER III, Ph.D. MIT 1985. An examination of depository institutions' funds management in a deregulated environment.

MARY K. BUMGARNER, Ph.D. Georgia State 1984. An examination of the municipal capital maintenance decision.

YANG-HO BYEON, Ph.D. Northern Illinois 1985. Effects of financial liberalization and monetary policy in the presence of unregulated money market.

WAYNE D. CARROLL, Ph.D. Minnesota 1984. The role of brokers in the creation of information-exchange networks.

CHIEN HSUN CHEN, Ph.D. Oklahoma State 1985. Impact of anticipated and unanticipated monetary and fiscal changes on output: An empirical study.

YEA-MOW CHEN, Ph.D. Ohio State 1984. Interest rate volatility, monetary policy, and inventory investment.

WU-JEN CHUANG, Ph.D. Indiana-Bloomington 1984. Inflation uncertainity and household flow of funds allocations.

JAMES A. CLOUSE, Ph.D. Northwestern 1985. Real balances and investment.

DEAN D. CROUSHORE, Ph.D. Ohio State 1984. Optimal government financial policy in a transactions-cost model of money.

THOMAS J. CUNNINGHAM, Ph.D. Columbia 1985. The asset side of money liabilities.

TIMOTHY DANIEL, Ph.D. Maryland 1984. Tax incidence in a two-sector model with three factors of production.

GEORGE K. DAVIS, Ph.D. Southern Methodist 1984. Determination of interest rates when inflation is uncertain.

MOHAMMAD K. DERAVI, Ph.D. Oklahoma State 1985. Income taxation and demand for money.

KATHLEEN K. DONAHOO, Ph.D. Yale 1984. Financial innovation in the savings and loan industry.

DEBRA ANN DRECNIK, Ph.D. Purdue 1984. Creative financing: A question of optimal housing finance.

KARL H. DUNZ, Ph.D. California-Berkeley 1984. Some results on economies with land: Equilibria and the core of economies with indivisible and divisible land.

PIERRE A. ELLIS, Ph.D. Chicago 1985. The role of borrower attributes in mortgage loan pricing.

SALEH M. EL-MANSOURI, Ph.D. Washington State 1984. Alternative money demand specifications: Their specifications and forecasting abilities, 1960:1 to 1982:3.

ANASTASSIOS GAGALES, Ph.D. Rochester 1985. The demand for money in Greece.

JOHN A. GILDEA, Ph.D. Duke 1985. A theory of federal open market committee voting behavior.

MARIA S. GOCHOCO, Ph.D. Columbia 1984. Anticipated money, unanticipated money, and interest rate movements in the United States: 1959–79.

KEVIN B. GRIER, Ph.D. Washington-St. Louis 1984. The political economy of monetary policy.

GIKAS HARDOUVELIS, Ph.D. California-Berkeley 1985. Three essays in policy regime changes, interest rate volatility, and Federal Reserve credibility.

JOSEPH G. HAUBRICH, Ph.D. Rochester 1985. Imperfect information and the role of banks.

EDWARD HJERPE III, Ph.D. Notre Dame 1985. Simultaneity and the demand for money.

DALYA INHABER, Ph.D. Michigan 1984. Anticipated fiscal and monetary policy switches.

MAHMOODUL ISLAM, Ph.D. Wayne State 1985. Short-term interest rates: A theoretical and empirical investigation.

MANSOR B. JUSOH, Ph.D. Washington State 1985. A model of the demand for money in moderate inflation.

ELIS W. KACAPYR, Ph.D. Georgia State 1985. The contribution of government bonds to net private wealth.

KWAN-SOO KIM, Ph.D. SUNY-Albany 1984. Credit rationing and nonprice terms on bank loans: An econometric approach using a switching equilibrium-disequilibrium model with price bounds.

JULES E. KING, Ph.D. Stanford 1985. Black savings and loan associations as distinct entitites of the savings and loan industry.

SUDHIR KRISHNAMURTHI, D.B.A. Harvard 1985. The impact of size on security return reaction to quarterly earnings announcements.

GREGORY A. KROHN, Ph.D. Wisconsin-Madison 1985. Interest rate volatility and the size of the financial seçtor.

EN-AI KUO, Ph.D. Oklahoma State 1985. Federal Reserve reaction functions and the effectiveness of monetary policy.

JEFFREY M. LACKER, Ph.D. Wisconsin-Madison 1984. Three essays in macroeconomics.

DENNIS J. LASSER, Ph.D. Indiana-Bloomington 1984. Hedging effectiveness and efficiency of Treasury bill and bond futures markets: Further evidence.

J. T. MARSHALL LEE, Ph.D. Queen's 1984. The effects of taxes on resource allocation and distribution of income in Canada: A general equilibrium analysis.

DAVID M. LEVERMORE, Ph.D. American 1984. The relevance of a secondary mortgage market institution to Jamaica.

JOSEPH LIM, Ph.D. Pennsylvania 1985. Testing monetarist and structuralist theories of inflation in the Philippines.

CHUAN LIN, Ph.D. Illinois-Urbana 1984. The incidence of the property tax.

WILLIAM A. LORD, Ph.D. Indiana-Bloomington 1984. Federal tax reform analyses in an overlapping-generations life cycle model with variable labor supply.

LARISSA S. LOZYNSKYJ, Ph.D. Columbia 1985. Inflation and nominal interest rates: Testing inflation predictions in the government securities market.

BERND P. LUEDECKE, Ph.D. Wisconsin-Madison 1984. An empirical investigation into arbitrage and approximate K-factor structure on large asset markets.

STEVEN A. LUMPKIN, Ph.D. Washington-St. Louis 1984. Rational bidding theory and U.S. Treasury bill auctions.

BASIL MANESSIOTIS, Ph.D. Northwestern 1985. Rates and monetary policy tax structure change during economic development: The case of Greece.

THOMAS H. MARKS, Ph.D. Georgia State 1985. A numerical approach to the incidence of the state corporate income tax.

PAUL D. MASTRODDI, Ph.D. Yale 1984. An examination of monetary policy rule.

DENNIS D. MILLER, Ph.D. Colorado-Boulder 1985. Interest rates as signals and the effects of their changes on relative prices.

ABDOL H. MOINI, Ph.D. Alabama 1985. The usefulness of credit aggregates as intermediate monetary policy targets.

RAHMAT MOZAYAN, Ph.D. West Virginia 1985. The impact of the redefined monetary aggregate on the structure of money demand in the United States.

AHMED H. NAGADI, Ph.D. Claremont 1985. The demand for money in Saudi Arabia: An empirical study.

STEPHEN D. OLINER, Ph.D. Wisconsin-Madison 1985. The effect of investment incentives on capital-use decisions: Theory and simulations.

GEORGE M. OLMSTEAD, Ph.D. Northwestern 1984. Tiebout and the demand for local public goods.

NITIN PARANJPE, Ph.D. Wayne State 1984. The impact of the change in the operating procedure of the Federal Reserve Board on interest rate levels and variability.

GONZALO C. PASTOR, Ph.D. Indiana-Bloomington 1985. Short-term control of monetary aggregates: The case of Peru.

MICHAEL T. PEDDLE, Ph.D. Northwestern 1985. Local tax differentials and the intrametropolitan location of firms: A simultaneous equations and panel analysis approach.

ALESSANDRO PENATI, Ph.D. Chicago 1984. Two essays on money, business cycles, and the real rate of interest.

DANIEL G. PERDUE, Ph.D. Alabama 1985. The portfolio approach to state and local tax structures.

HASSAN PIRASTEH, Ph.D. Oregon 1985. An econometric analysis of Oregon state fiscal response to federal grants.

AKELLA S. RAO, Ph.D. Northwestern 1984. Ownership structure and expense preference in the savings and loan sector: Theory and evidence.

RAHGOZAR G. REZA, Ph.D. Claremont 1985. Personal saving and inflation uncertainty.

JUDITH A. ROBERTS, Ph.D. Michigan 1985. Housing tenure and demand for local public goods.

ANTONIO J. RODRIGUEZ, Ph.D. Alabama 1985. Futures trading and cash market rates variability: An empirical study of the Treasury bill markets.

DAVID ROMER, Ph.D. MIT 1985. General equilibrium analyses of government financial policies.

BENJAMIN RUSSO, Ph.D. Iowa 1985. Inflation uncertainty, demand phenomenal, assets, and interest rates.

DONALD SABBARESE, Ph.D. Georgia State 1984. A market structure-conduct-performance model for banking.

JON W. SALMON, Ph.D. California-Los Angeles 1985. The emergence of organized futures markets: The distribution of consumption risk.

DANIEL SCHWALLIE, Ph.D. Iowa 1984. Intergovernmental grants and the aggregate public sector.

AARON SCHWARTZMAN, Ph.D. MIT 1985. Exchange rate and the tax structure in a *Q* model of investment: The Mexican experience.

APOSTOLOS SERLETIS, Ph.D. McMaster 1984. The substitutability and separability of monetary assets.

MOJTABA SEYEDIAN, Ph.D. SUNY-Binghamton 1984. Wealth effects of corporate bond valuation on consumption.

REED H. SHULDINER, Ph.D. MIT 1985. Markets in taxes.

GERASIMOS SOLDATOS, Ph.D. Clark 1985. Income vs. consumption base taxes: A multiperiod analysis of equity.

EDMUND NII KWAKU SOWA, Ph.D. McMaster 1985. Multiproduct cost modelling of Canadian trust companies.

PETER STELLA, Ph.D. Stanford 1985. Credibility and the effectiveness of macroeconomic policy.

SCOTT B. SUMNER, Ph.D. Chicago 1985. An empirical study of currency hoarding.

DAVID D. VANHOOSE, Ph.D. North Carolina-Chapel Hill 1984. Bank market structure and monetary policy.

TSONG-MIN WU, Ph.D. Rochester 1984. Optimal taxation over time and the time inconsistency problem.

PAUL G. WYCKOFF, Ph.D. Michigan 1984. Bureaucracy, efficiency, and local public choice: An empirical test of some new implications of bureaucratic public service provision.

DAVID CHIE-HWA YANG, Ph.D. Columbia 1985. The association between SFAS 33 information and bond ratings.

PYUNG I. YU, Ph.D. Catholic 1985. The impact of the change of the Korean indirect tax system from turnover and excise taxes to value-added tax on the income distribution.

KHAN H. ZAHID, Ph.D. Columbia 1985. Monetary accommodation of government budget deficits.

MARK P. ZAPOROWSKI, Ph.D. SUNY-Albany 1985. Inflation expectations and the rate of interest: An econometric analysis.

International Economics

NISO ABUAF, Ph.D. Chicago 1985. The effects of monetary and real disturbances on the real exchange rate.

TAWANG ALUN, Ph.D. Vanderbilt 1984. An economic analysis of developing countries' external indebtedness.

GHODRATOLLAH ARABIAN, Ph.D. Indiana-Bloomington 1985. The impact of exchange rate uncertainty on exports of developing countries: Case of Korea.

ERIC N. K. ASANTE, Ph.D. McMaster 1985. A mathematical programming model of trade and protection applied to the Canadian textile sector.

ABDULAZIZ A. S. AS-SUDAIS, Ph.D. Colorado-Boulder 1985. The applicability of the theory of customs union to the case of the Gulf Cooperation Council.

SAYEEDA BANO, Ph.D. Simon Fraser 1985. The extent and determinants of intra-industry international trade in Canada: An empirical investigation.

LUCA BARBONE, Ph.D. MIT 1985. Essays on trade and macro policies in developing countries.

LLOYD D. BARNEY, JR., Ph.D. Texas A&M 1984. A theoretical analysis of international trade with exhaustible natural resources.

DRUSILLA K. BROWN, Ph.D. Michigan 1984. A simulation model of world production and trade with bilateral trade and commercial policies and application to the Tokyo round.

EUI TAE CHANG, Ph.D. Wisconsin-Madison 1984. Adjustment of quality mix to quantitative restrictions and protective effects.

MARTIN CHERKES, Ph.D. Pennsylvania 1984. The theory of the multinational enterprise in manufacturing industries.

JANG BONG CHOI, Ph.D. Pittsburgh 1984. The stochastic behavior of spot and forward exchange rates.

SANG-RIM CHOI, Ph.D. Michigan 1984. Economic rational and consequences of international banking and financial centers.

HAN G. CHOO, Ph.D. Columbia 1985. Three essays in open economy macroeconomics.

CHARLES J. CORRADO, Ph.D. SUNY-Albany 1985. Official foreign exchange market intervention by the industrialized countries in the 1970's.

LUIS C. CUNHA, Ph.D. Columbia 1985. Transport costs and international trade: Issues and consequences.

FUMIO DEI, Ph.D. Rochester 1985. Essays on foreign investment.

HARRIS DELLAS, Ph.D. Rochester 1985. Cyclical comovements in real economic activity and prices in the world economy.

MARYANN F. DILIBERTO, Ph.D. Columbia 1984. A test for structural change in U.S. trade and international capital investment with major industrial countries.

ELIAS DINOPOULOS, Ph.D. Columbia 1985. Essays on trade and welfare.

ISAAC DIWAN, Ph.D. California-Berkeley 1984. Welfare aspects of international trade in goods and in securities.

MICHAEL A. DUFFY, Ph.D. Chicago 1985. The exchange rate relative prices and the terms of trade: A dynamic asset market approach.

JOE FALZON, Ph.D. Northwestern 1984. Exchange rate policy, monetary policy, and wage indexation within the European Economic Community: A theoretical analysis.

LAUREN FEINSTONE, Ph.D. Chicago 1985. Intradaily market efficiency and price processes in the futures market in foreign exchange.

THOMAS A. FETHERSTON, Ph.D. Rutgers-New Brunswick 1985. Risk compensation in Eurolending: An analysis of the risk-return tradeoff in Eurolending for the period 1975 through 1981.

ERIC FISHER, Ph.D. California-Berkeley 1985. Three essays in international trade theory.

EDWARD L. FITZSIMMONS, Ph.D. Nebraska-Lincoln 1984. The Pacific Coast export market for corn: Past growth and future prospects.

ARMINIO FRAGA, Ph.D. Princeton 1985. Essays in international lending and adjustment.

KWOK-CHIU FUNG, Ph.D. Wisconsin-Madison 1984. International rivalry, intra-industry trade, and the effects of trade policies.

JOHN R. GARRETT, Ph.D. Massachusetts-Amherst 1985. The post-World War I international financial order.

PAUL R. GOLDING, Ph.D. Fletcher School 1984. Effectiveness of escape clause relief in promoting adjustment to import competition.

DANIEL GROS, Ph.D. Chicago 1984. Increasing returns and human capital in international trade.

VAIKUNTA M. GULIVINDALA, Ph.D. New School 1985. A critical examination of LDC's debt from official sources and private sources: Case studies of Brazil, Mexico, India, and Thailand.

NOBORU HAMAGUCHI, Ph.D. Michigan 1985. Structural change in Japanese-American interdependence: A total factor productivity analysis in an international input-output framework.

DAVID E. HANIG, Ph.D. George Washington 1985. The Heckscher-Ohlin theorem and Spanish foreign trade.

C. RANDALL HENNING, Ph.D. Fletcher School 1985. The politics of macroeconomic conflict and coordination among advanced capitalist states, 1975–85.

AWANG HUSSIN, Ph.D. Pennsylvania 1984. Aspects of financial policy in an open developing economy: The Malaysian example.

PAUL JOHNSON, Ph.D. Rochester 1985. Essays in international economics.

MIHAIL KABALOURIS, Ph.D. McMaster 1985. Purchasing power parity and exchange rate overshooting: An econometric investigation.

HO SANG KANG, Ph.D. Columbia 1984. A model of temporary deviations from purchasing power parity.

KRISHNA M. KASIBHATLA, Ph.D. Rutgers-New Brunswick 1984. Factors behind the Brazilian external debt crisis, and an alternative plan to large-scale rescheduling.

MOHAMMED N. G. KHAYAT, Ph.D. Claremont 1985. A portfolio approach to the oil surplus problem: Saudi Arabia as a study case.

WI-CHAN KIM, Ph.D. Michigan 1984. The emerging patterns of industrialization, intermediate product trade, and foreign direct investment in the east Pacific rim.

RICHARD W. KJETSAA, Ph.D. Fordham 1985. An empirical examination of international monetarism: A case study of Sweden.

THOMAS KLITGAARD II, Ph.D. Stanford 1985. Informative exchange rate movements and the U.S. macroeconomy.

LINDA KOLE, Ph.D. MIT 1984. Essays on stabilization policies under flexible exchange rates.

KALA KRISHNA, Ph.D. Princeton 1984. Trade restrictions under imperfect competition.

TAE WOON KWAK, Ph.D. Ohio State 1985. International transmission of economic disturbances under floating exchange rates.

RALPH G. LANGENBORG Ph.D. California-Berkeley 1985. The deutschemark 1973–82: A portfolio balance approach.

PATRICIA M. LASONDE, Ph.D. Boston College 1984. The

Heckscher-Ohlin therom and the factor content of trade: Empirical evidence.

HAHN KOO LEE, Ph.D. Kansas State 1984. A study of foreign exchange futures markets.

KANG WON LEE, Ph.D. Johns Hopkins 1985. Budget financing and monetary control in an open economy: Implications for sterilized intervention under fixed exchange rates.

KI SEOK LEE, Ph.D. Rutgers-New Brunswick 1985. Shifting patterns of international trade: The United States, Japan, and Korea.

KAREN K. LEWIS, Ph.D. Chicago 1985. The effectiveness of sterlized intervention through a portfolio balance channel: An empirical investigation.

PETER LIRITZIS, Ph.D. Clark 1985. Shadow prices for project evaluation—A World Bank approach for Greece.

BIH JANE LIU, Ph.D. Wisconsin-Madison 1985. Trade and tax policies in distorted small economies.

MICHAEL MERED, Ph.D. George Washington 1985. Exchange rate variability: The Korean experience in a world of floating rates.

ROBERT METTS, Ph.D. California-Berkeley 1985. Legislative aspect of international trade: The case of Kenya colony and protectorate, 1895–1963.

KAZUYUKI MIYAGIWA, Ph.D. Texas-Austin 1985. Essays on international capital mobility and commercial policies.

DANIEL A. MIZAK, Ph.D. Pittsburgh 1984. Economic growth and external debt: An analysis of the effects of parameter variability and interaction.

ADWAIT K. MOHANTY, Ph.D. Wisconsin-Milwaukee 1984. Welfare and policy implications of dumping from the exporting country's point of view.

ROGER V. MOORE, Ph.D. Rutgers-New Brunswick 1985. A growth model of the Eurodollar market.

JOHN J. MUSSACHIA, Ph.D. California-Santa Barbara 1984. A reexamination of the PPP doctrine during the recent floating rate period.

LUIZ P. NUNES-MARINHO, Ph.D. Chicago 1984. Optimal capital accumulation and external indebtedness in a two-sector small economy model.

TONG-WHEE OH, Ph.D. Indiana-Bloomington 1985. Export-oriented exchange rate policy, foreign capital inflows, and sterilization: The case of Korea.

DAVID R. ORDEN, Ph.D. Minnesota 1984. Transfers, exchange rates, and agricultural markets.

MILTON PAPPAS, Ph.D. Johns Hopkins 1985. A q theoretic approach to economic growth, exchange rate dynamics, and the balance of international indebtedness.

WON AM PARK, Ph.D. MIT 1985. Three essays on stabilization and speculation.

MANUEL PASTOR, JR., Ph.D. Massachusetts-Amherst 1984. The international monetary fund and Latin America: Core-periphery and capital-labor aspects of fund policy.

MARIO H. PASTORE, Ph.D. American 1984. Export instability, international business cycles, and commodity concentration of LDCs.

SHULAMIT PESSACH, Ph.D. Chicago 1984. Essays on exchange rate determination and capital controls.

ROBERT L. POWELL, Ph.D. California-Berkeley 1985. Nuclear deterrance theory; trade and a security dilemma: Uncertainty and interest elasticity.

DANI RODRIK, Ph.D. Princeton 1985. Studies on the welfare theory of trade and exchange rate policy.

ALEJANDRO SANTOS-LEAL, Ph.D. Yale 1985. Three essays on the nominal and real exchange rate in Mexico: 1925–81.

ANTOINETTE M. SAYEH, Ph.D. Fletcher School 1985. Monetary dependence, payments disequilibria, and fiscal policy in a small, open, low-income economy: Experiments with a hybrid model of Liberia.

LAWRENCE L. SCHEMBRI, Ph.D. MIT 1985. Essays on the economic interdependence of Canada and the United States.

ARDESHIR SEPEHRI, Ph.D. Alberta 1985. A monetary model of the balance of payments: Tanzania, 1962–81.

PETER S. SEPHTON, Ph.D. Queen's 1984. Foreign shocks and exchange market intervention in a two-country model: Evidence from Canada and the United States.

PABLO J. SERRA, Ph.D. Yale 1984. Openness and national advantage in less developed countries.

SE DON SHIN, Ph.D. California-Los Angeles 1984. A theory of exchange rate, relative price of commodity, and real wage in a neoclassical model of accumulation.

RONALD L. SOLBERG, Ph.D. California-Berkeley 1984. Sovereign risk, portfolio management, and debt rescheduling.

ANDREW SOLOCHA, Ph.D. Michigan State 1985. Classical comparative cost theory: Time-series and cross-sectional evidence for the United States, Japan, and Europe.

MOHAMED O. SULIMAN, Ph.D. Indiana-Bloomington 1984. Appropriate exchange regimes and trade structure for small, open, developing economies in the context of trade controls.

JOHN M. TOBIN, Ph.D. Fordham 1985. An application of the arbitrage theory of capital asset pricing to the syndicated Eurocurrency credit.

LILY V. UY, Ph.D. George Washington 1984. Creditworthiness of less developed countries: An analysis of factors affecting lenders' perceptions.

SUSHIL VACHANI, D.B.A. Harvard 1985. Strategic product market choices by multinationals and local firms in a newly industrializing country.

MICHAEL VAN BEMMELEN, Ph.D. Columbia 1985. Some theoretical and empirical aspects to the international mobility of factors of production.

STEVEN VAUGHAN, Ph.D. John Hopkins 1985. Exchange rate determination and the trade balance with leads and lags.

ALEJANDRO VIOLANTE-MORLOCK, Ph.D. American 1985. Trade liberalization between Mexico and the United States.

TUN A. WAI, Ph.D. Georgetown 1984. Dynamic credit rationing in the Eurodollar market: Case of Brazil.

THOMAS WEBSTER, Ph.D. City-New York 1985. Analyzing country risk: Predicting the probability of external debt repudiation in the post-oil embargo decade.

HARRY L. WHITE, Ph.D. Texas A&M 1984. Exchange rate risk and the multinational firm.

EDY L. WONG, Ph.D. Alberta 1984. An empirical study of the interplay between the Asian currency market and the Asean economies.

CHUNG-SHU WU, Ph.D. Northwestern 1984. Exchange rates, interest rates, and monetary policy.

SHMUEL YAHALOM, Ph.D. City-New York 1984. Exchange rates and inflation in Israel.

HOSSEIN N. YARANDI, Ph.D. Indiana-Bloomington 1984. The monetary approach to the determination of effective exchange rates in small open economies.

YIN BERNARD YEUNG, Ph.D. Chicago 1984. Exchange rate intervention policies, rational expectations, interdependence of policy actions, and policy expectations.

PHILIP M. YOUNG, Ph.D. California-Berkeley 1984. Uncertainty, the cost of hedging exchange risk, and international trade.

JAIME ZABLUDOWSKY, Ph.D. Yale 1984. Money, foreign indebtedness, and export performance in Porfirist Mexico.

IQBAL M. ZAIDI, Ph.D. Princeton 1984. Macroeconomic adjustments and the balance of payments in developing countries.

KAREL ZELENKA, Ph.D. Fletcher School 1985. U.S. export controls in a business perspective.

Business Administration; including Business Finance and Investment, Insurance Marketing, and Accounting

SHIRAZ T. ADAM, Ph.D. Simon Fraser 1984. Predictive ability of annual cash flow under historical cost and constant dollar accounting.

KENNETH D. ARMSTRONG, Ph.D. Northwestern 1985. The effect of power on the role of the governing board in the governance of the private colleges and universities of Illinois.

SONDHI C. ASHWINPAUL, Ph.D. New York 1985. Finance subsidiaries and debt capacity.

AJAY BHASIN, Ph.D. Northwestern 1985. The effects of memory organization on product judgments: Comparison vs. category strategy.

DEBORAH G. BLACK, Ph.D. New York 1985. Success and failure of futures contracts: Theory and empirical evidence.

PAUL M. BOBROWSKI, Ph.D. Indiana-Bloomington 1985. An investigation of alternate routing in the job shop as a planning process.

JOHN E. BUTLER, Ph.D. New York 1985. Opportunity and entrepreneurship: Strategic links in the competitive process.

JAY A. CONGER, D.B.A. Harvard 1985. Charismatic leadership in business: An exploratory study.

ANTHONY D. COX, Ph.D. Indiana-Bloomington 1984. A study of the merchandise sales projections of department store buyers.

JOHN DUMOND, Ph.D. Indiana-Bloomington 1984. Evaluation of due date assignment rules for dynamically arriving projects.

BARBARA B. FLYNN, D.B.A. Indiana-Bloomington 1984. Group technology layout vs. process layout: A comparison using job shop simulation as the means for multiple attribute comparison.

E. JAMES FLYNN, Ph.D. Indiana-Bloomington 1985. The effects of decline and dependence on administrative components.

WEE-BENG GAN, Ph.D. Pennsylvania 1984. An economic analysis of the financial balance sheet of the industrial and commercial companies sector in the United Kingdom.

GULIZ GER, Ph.D. Northwestern 1985. Affect and judgement.

KATHLEEN M. HAGERTY, Ph.D. Stanford 1985. Equilibrium bid-ask spreads in a dealership market.

KAREN HARTMAN, Ph.D. New York 1985. A framework of routine learning: An experimental study of the learning of a computer-based task.

MICHAEL HARTZMARK, Ph.D. Chicago 1984. The distribution of large trader returns in futures markets: Theory and evidence.

JOHN R. HOLLENBECK, Ph.D. New York 1984. Control theory, self-focus and behavior in organization.

DORIS L. HOLT, Ph.D. Michigan 1984. Evidence integration in the formation of risk assessments by auditors and bank lending officers.

STEVEN ISBERG, Ph.D. SUNY-Binghamton 1984. Incremental information and security returns: A theoretical discussion with an empirical application to corporate dividend policy.

YOKO ISHIKURA, D.B.A Harvard 1985. Global new product development: The case of plain paper copiers.

JANICE M. K. JAFERIAN, D.B.A. Harvard 1985. Decision management: A study of the interaction between the formal and informal executive decision-making processes.

RICHARD D. JUDY, Ph.D. Indiana-Bloomington 1984. Interorganizational organizational linking mechanisms: Health systems agencies as a case in point.

PATRICK J. KAUFMANN, Ph.D. Northwestern 1985. Relational exchange and conflict behavior: An investigation of their relationship to hostility in the conflict aftermath.

GARY M. KERN, Ph.D. Indiana-Bloomington 1985. An investigation of the interaction between police patrol allocation and dispatching policy.

CHANKON KIM, Ph.D. Indiana-Bloomington 1985. A study of husband-wife decision making in a conflictual situation.

HYUK KIM, Ph.D. Indiana-Bloomington 1985. An empirical study of the market reaction to Statement of Financial Accounting Standards No. 52: Foreign currency translation.

DANA F. KLEMME, Ph.D. Michigan 1984. A study of line-of-business disclosures: Accounting vs. market risk measures.

GERARD P. LEARMONTH, Ph.D. Michigan 1983. Adaptive data management.

HANJOON LEE, Ph.D. Indiana-Bloomington 1984. Factors affecting the use and evaluation of marketing information: An experimental approach.

LAURA A. LEHNER, Ph.D. SUNY-Albany 1985. Determinants of the cost of capital for electric utilities.

GEORGE H. LENTZ, D.B.A. Indiana-Bloomington 1985. A comparison of the impact of various tax law provisions on depreciable assets relative to tax-neutral models.

DANIEL A. LEVINTHAL, Ph.D. Stanford 1985. Three essays on economic models of organizations.

PU LIU, Ph.D. Indiana-Bloomington 1985. The information content in bond ratings and potential bond rating changes.

URI LOEWENSTEIN, Ph.D. New York 1984. The effects of risk on dividend policy: Related essays.

MADGE LYMAN, Ph.D. Columbia 1985. Sales interaction and buyer behavioral model.

NANCY R. MANGOLD, Ph.D. California-Berkeley 1984. The effect of auditor changes on earnings, opinion, and stock prices.

PARVIZ MANOUCHEHRI-ADIB, Ph.D. Texas-Austin 1985. An investment planning model of the world petrochemical industry.

DAVID J. MOORE, Ph.D. Indiana-Bloomington 1984. Consumer response to price discounts as a function of brand name and store type: An information-processing explanation.

KHALED A. MUTTAR, Ph.D. West Virginia 1985. An investigation of the validity of objective and subjective measures of organizational climate.

HARRY A. NEWMAN, Ph.D. Northwestern 1984. Capital investment, managerial incentives, and corporate taxes.

JANE AI-JIUAN OU, Ph.D. California-Berkeley 1984. The information content of nonearnings accounting numbers as earnings predictors.

PAUL-VICTOR PARE, Ph.D. Michigan 1983. An empirical evaluation of composite predictors of accounting data.

RODNEY A. PEARSON, D.B.A. Harvard 1985. The role of information systems in today's railroad.

MEYER J. PEIKES, Ph.D. Columbia 1985. Economic, financial, and mangerial implications of the incentive signalling model: A two-variable approach.

PAMELA A. POSEY, D.B.A. Harvard 1985. Excellence in first line supervision: A comparative study of traditional and innovative work systems.

S. E. C. PURVIS, Ph.D. Columbia 1985. The auditor's preliminary evaluation of internal accounting control: A behavioral analysis.

ANTONIO L. QUE, Ph.D. New York 1985. A critical evaluation of FASB Statement No. 13.

GARY L. RAGATZ, Ph.D. Indiana-Bloomington 1985. An evaluation of order release mechanisms for job shops.

ARVIND RANGASWAMY, Ph.D. Northwestern 1985. Sales force organization: A modeling approach.

JAMES E. REBELE, Ph.D. Indiana-Bloomington 1984. An investigation of the impact of source expertise in the auditor-client relationship.

WILLIAM T. ROBINSON, Ph.D. Michigan 1984. Market pioneering and market share in consumer goods industries.

GARY J. RUSSELL, Ph.D. Chicago 1985. The impact of decision rules on advertising measurement.

C. M. SASHI, Ph.D. Northwestern 1985. Product differentiation and market performance in producer goods industries.

JOHN D. SCHATZBERG, Ph.D. Michigan 1984. Voluntary liquidation and stockholder returns.

JOAN T. SCHMIT, Ph.D. Indiana-Bloomington 1984. Erosion of the workers' compensation exclusivity rule: A consideration of employers' dual capacity as product manufacturers.

SYED S. A. SHAH, Ph.D. Indiana-Bloomington 1984. Asymmetric information and contractual arrangements: Applications to asset management and project financing.

HARBIR SINGH, Ph.D. Michigan 1984. Corporate acquisitions and economic performance.

RAVI SINGH, Ph.D. Northwestern 1985. The environment as an enacted perceptual field: A theoretical, methodological, and empirical investigation in the context of marketing channels.

GILBERT L. SKILLMAN, JR., Ph.D. Michigan 1985. Microfoundations of the conflict theory of the firm.

BADIH C. SOUBRA, Ph.D. New York 1985. Optimality of debt covenants.

JAMES SOWERS, Ph.D. Columbia 1984. Control in organizations: An empirical test of the consistency model of control.

RALPH E. STABLEIN, Ph.D. Northwestern 1985. Why does goal setting work? An examination of the motivational antecedents of the goal-performance relationship.

GERARD J. TELLIS, Ph.D. Michigan 1983. Consumer information, marketing strategy, and market share.

JAMES B. THOMSON, Ph.D. Ohio State 1984. Tax-selling pressure and errors in recorded security prices: An empirical investigation of the turn-of-the-year effect.

NICKOLAOS G. TRAVLOS, Ph.D. New York 1984. Takeovers and method of payment: The effect of alternative financing schemes on common stock prices of bidding and target firms.

JAMES P. WALSH, Ph.D. Northwestern 1985. The schema for organizational success: Its origins, structure, and implications for strategic decision making.

WILLIAM A. WEEKS, D.B.A. Indiana-Bloomington 1984. Improved measurement for lifestyle concept: A "time" segmentation approach.

PAUL A. ZAROWIN, Ph.D. Chicago 1985. Empirical evidence on the tax effects of inflation.

Industrial Organization and Public Policy; including Economics of Technological Change, and Industry Studies

DEAN AMEL, Ph.D. MIT 1984. Research and development with competing innovations: Theoretical implications and empirical evidence.

SUE A. AUGUSTYNIAK, Ph.D. Michigan 1984. A multinomial logit analysis of the choice of transportation mode.

HYUNG BAE, Ph.D. Wisconsin-Madison 1985. Three essays in industrial organization theory.

SRINIVASAN BALAKRISHNAN, Ph.D. Michigan 1983. A dynamic analysis of the vertical integration strategy.

SUZANNE BARTMAN, Ph.D. Wayne State 1984. A disaggregate approach to the modeling of automobile demand: Determinants of the level and the rate of change in Dode and Hamilton counties.

SIDNEY L. BARTON, Ph.D. Indiana-Bloomington 1985. The corporate strategy-capital structure relationship in the large firm.

RANDALL W. BENNETT, Ph.D. Michigan State 1984. Quality of service in nonmonopoly air travel markets: An empirical investigation of measures of service quality.

GREGORY A. BISCHAK, Ph.D. New School 1985. State induced technological change in the U.S. nuclear power industry, 1947-83: A case study of the interaction of complementary public investment and private development.

CARLOS A. BRAGA, Ph.D. Illinois-Urbana 1984. Steel in a newly industrialized country: The case of Brazil.

MICHAEL E. BURTON, Ph.D. California-Los Angeles 1985. The 1901 establishment of the U.S. Steel Corporation: For monopoly and/or efficiency?

ROBIN A. CANTOR, Ph.D. Duke 1985. An analysis of public costs and risks in the Canadian nuclear industry.

JEFFREY L. CARLSON, Ph.D. Illinois-Urbana 1984. The law and economics of hazardous waste landfill disposal: A framework of analysis.

PHILIP C. CHANG, Ph.D. Illinois-Urbana 1985. Conglomerate mergers and the pursuit of economies of scope.

CHIA-YON CHEN, Ph.D. West Virginia 1984. Impact analysis and optimal adjustment pattern during a supply restriction in aluminum: A case study for Taiwan.

THOMAS A. CHRISTIAN, Ph.D. Columbia 1985. An examination of selected topics related to mergers and aquisitions.

THOMAS E. CLIFFORD, Ph.D. California-Santa Barbara 1984. The external costs of electric power for California from coal-fired and nuclear power plants.

THOMAS E. COOPER, Ph.D. Princeton 1984. Facilitating practices and most-favored-customer pricing.

WALTER D. CRUICKSHANK, Ph.D. Pennsylvania State 1985. The determinants of the restructuring of the U.S. zinc smelting industry.

DONNA L. DOANE, Ph.D. Yale 1984. Two essays on technological innovation: Innovation and economic stagnation, and interfirm cooperation for innovation in Japan.

MATTHEW GELFAND, Ph.D. Pennsylvania 1984. Empirical issues in industrial organization: The case of the Uruguyan banking sector.

JAMES T. GIKAS, Ph.D. California-Los Angeles 1984. Trading stamps and price discrimination.

THOMAS W. GILLIGAN, Ph.D. Washington-St. Louis 1984. Product distribution with vertical price control.

MARK GLICK, Ph.D. New School 1985. The theory of industrial pricing: An empirical analysis of profit rate dispersion.

ROBERT C. GRAHAM, Ph.D. Illinois-Urbana 1985. Diffusion of technology: A dynamic approach.

ELIZABETH HALL, Ph.D. Boston College 1985. An analysis of preemptive capital expansion and the titanium dioxide industry.

JAMES A. HARRIS, SR., Ph.D. Pacific Western 1983. International industrial organization economics of the multinational banking system.

THOMAS W. HAZLETT, Ph.D. California-Los Angeles 1984. Three essays on monopoly.

CONSTANCE HELFAT, Ph.D. Yale 1985. Oil industry investment and research as portfolio choices.

LEN-KUO HU, Ph.D. California-Los Angeles 1984. Monopolistic competition, product differentiation, and intra-industry trade.

ALBERT W. ISENMAN, Ph.D. Northwestern 1984. The relationship of relative power and business strategy to business unit performance.

ANDREW S. JOSKOW, Ph.D. Yale 1984. A welfare analysis of industry product variety: The automobile industry.

IOANNIS N. KESSIDES, Ph.D. Princeton 1985. Empirical tests of oligopoly hypotheses.

CHI-MING KOO, Ph.D. New York 1985. The economic impact of changing automotive materials use.

HEIDI KROLL, Ph.D. California-Berkeley 1985. Law and plan: An economic analysis of contractual relations in Soviet industry.

KISANG LEE, Ph.D. Northwestern 1984. Optimal timing of international technology transfer in the presence of technology diffusion effects.

JOSEPH A. LEVIN, Ph.D. Boston College 1985. The potential entrant and the decision about entry: Two case studies.

ROBERT G. LYNCH, Ph.D. SUNY-Stony Brook 1984. The Catholic Church as a resource allocation mechanism.

DONALD L. MCCABE, Ph.D. New York 1985. The impact of environmental uncertainty on buying group structure and effectiveness.

KAI MA, Ph.D. Iowa 1984. Theory of the firm: A bargaining approach.

DAVID W. MARTIN, Ph.D. Illinois-Urbana 1984. Spent nuclear fuel disposal externalities.

J. O. MATTHEWS, Ph.D. Temple 1985. The securities industry: An industrial organization analysis.

PHILLIP J. MIZZI, Ph.D. Texas A&M 1984. A dominant firm in an exhaustible resource industry: The case of nickel.

SCOTT E. MOLIN, Ph.D. New School 1985. The origin of the machinery industry in the United States: A regional history of mechanization in the nineteenth century.

JOHN H. MULHERIN III, Ph.D. California-Los Angeles 1984. Vertical integration and long-term contracts in the natural gas industry.

MICHAEL C. MUNGER, Ph.D. Washington-St. Louis 1984. Institutions and outcomes: Two essays on the importance of legislative structure for understanding public policy.

JOSEPH G. NAGY, Ph.D. Minnesota 1984. The Pakistan agricultural development model: An economic evaluation of agricultural research and extension expenditures.

MICHAEL C. NAUGHTON, Ph.D. Boston College 1984. The efficiency and equity of electricity price regulation: A two-part framework.

EDGAR A. NORTON, Ph.D. Illinois-Urbana 1984. Diversification and economies of scope: A theory of the multiproduct firm.

IAN NOVOS, Ph.D. Pennsylvania 1985. Essays in applied microeconomics.

DOUGLAS R. OSTROM, Ph.D. Michigan 1984. Postwar Japanese industrial policy and changes in industrial structure.

JAESUN PARK, Ph.D. Northwestern 1985. Innovation competition under uncertainty.

JAMES A. PARROTT, Ph.D. Massachusetts-Amherst 1985. Technological and institutional innovation in Massachusetts electronics.

DONIS N. PETERSAN, Ph.D. Nebraska-Lincoln 1984. Residential electricity demand and optimal power resource utilization.

OLAYINKA PHILLIPS, Ph.D. Indiana-Bloomington 1985. Some regulatory aspects of domestic air passenger transportation in Nigeria: A policy analysis.

MOHAMMED RAFIQUZZAMAN, Ph.D. Carleton 1985. Market structure, risk, and the optimal patent term.

G. RAGHURAM, Ph.D. Northwestern 1984. Rail freight operations and the classification policy: Models and solution procedures.

SYLVIA RAW, Ph.D. Massachusetts-Amherst 1985. The political economy of Brazilian state-owned enterprises: 1964–80.

GARY ROBERTS, Ph.D. Pennsylvania 1984. Welfare maximization and competitive price discrimination.

DAVID I. ROSENBAUM, Ph.D. Wisconsin-Madison 1985. A study of firm strategies designed to deter entry as revealed in three antitrust cases.

EDWIN ROSENBERG, Ph.D. North Carolina State 1985. Input choice under uncertainty: An application of the capital-asset pricing model to the electric utility industry.

ARTHUR ROSENZWEIG, Ph.D. New York 1985. Quantity discounts in input markets and the cost function.

DONALD RYAN, Ph.D. SUNY-Albany 1984. Electricity demand: Estimating the impact of a nonexperimental time-of-use rate on load curves.

DAVID C. SALMON, Ph.D. Minnesota 1984. An evaluation of investment in agricultural research in Indonesia, 1965–77.

C. M. SASHI, Ph.D. Northwestern 1985. Product differentiation and market performance in producer goods industries.

CARL A. SCHERAGA, Ph.D. Connecticut 1985. The effect of regulation on the rate of adoption of cost-saving scale technology in the electric utility industry: A portfolio approach.

MARK J. SCHLESINGER Ph.D. Wisconsin-Madison 1984. Public, for profit, and private nonprofit enterprises: A study of mixed industries.

LOUIS SCHORSCH, Ph.D. American 1985. Demand factors import restraints and the prospects of the U.S. steel industry.

ABDUL H. SHAIKH, Ph.D. Boston 1985. Efficiency in production and public ownership: Theory and empirical analysis.

LYNN M. SHISHIDO-TOPEL, Ph.D. California-Los Angeles 1984. An economic analysis of exclusive dealing arrangements.

C. BRADLEY SHRADER, Ph.D. Indiana-Bloomington 1984. An investigation of the relationships among organization context, structure, and internal network properties.

MERRILE SING, Ph.D. Wisconsin-Madison 1984. The cost structure of the gas distribution and electricity industries.

ANNE E. SMITH, Ph.D. Stanford 1984. Market structure and regulatory design.

DONNA M. SMITH, Ph.D. Pennsylvania State 1984. The effect of economic regulation on risk and financial performance in the electric utility industry.

TIMOTHY R. SMITH, Ph.D. Washington 1984. Modeling an extractive industry: An application to the Canadian metal mining industry.

CYNTHIA L. SPANHEL, Ph.D. Indiana-Bloomington 1984. Firm capabilities as influences on the internalization vs. licensing decision.

EDWARD A. SYNDER, Ph.D. Chicago 1984. Defensive effort and efficient enforcement: An application to antitrust.

MARK A. STEPHENS, Ph.D. Tennessee-Knoxville 1985. An analysis of the Brazilian motor vehicle industry: The political economy of its development and its significance to the Brazilian economy.

NIKOLAOS A. STERGIOULAS, Ph.D. New Mexico 1985. The structure of Greek manufacturing and its demand for labor: An empirical study, 1966–76.

PATRICK A. TAYLOR, Ph.D. Alabama 1985. An estimation of the value of landing slots at four congested airports.

IGOR TOMIC, Ph.D. City-New York 1985. Electrical equipment purchasing patterns of utilities and commercial companies as affected by collusion among equipment manufacturers: An aspect of rate-base padding by electrical utilities.

PRAJAPATI TRIVEDI, Ph.D. Boston 1985. Comparative performance evaluation and explanation of public and private enterprise behavior.

VRAJAINDRA B. UPADHYAY, Ph.D. McMaster 1985. A reassessment of the Indian heavy industrialization strategy.

RICHARD WALL, Ph.D. SUNY-Buffalo 1984. Efficient capital markets and rate of return regulation.

DONG-HUA WANG, Ph.D. SUNY-Buffalo 1985. Economics of declining industry: American steel.

DONALD L. WESTERFIELD, Ph.D. Saint Louis 1984. Ananalysis of the divestiture of AT&T with regard to antitrust and the public interest.

KEVIN WYNN, Ph.D. Fordham 1985. The decision of electric utilities to complete or cancel nuclear reactors.

JOYCE M. YANCHAR, Ph.D. Boston College 1984. The demand for energy in the U.S. residential sector.

DENNIS A. YAO, Ph.D. Stanford 1984. Three essays on regulation with application to the automobile industry.

Agriculture and Natural Resources

ERNUR D. ABAAN, Ph.D. Michigan State 1985. The impact of base price policy in the agricultural sector: The Turkish experience.

ADESOJI O. ADELAJA, Ph.D. West Virginia 1985. An analysis of production technology and policy in the multiproduct cost function framework: The case of West Virginia agriculture.

AHMAD AFRASIABI, Ph.D. West Virginia 1985. A frequency domain of copper, lead, and zinc markets.

MOHAMMED A. ALSAHLAWI, Ph.D. Wisconsin-Milwaukee 1985. Spatial analysis of world energy markets: Demand for energy estimation and variable supply simulation.

JULIAN M. ALSTON, Ph.D. North Carolina State 1984. An analysis of international land prices.

PERVAIZ AMIR, Ph.D. Michigan State 1985. Increasing the productivity of rice-based farming systems of the Pakistan Punjab.

ERIC E. ANDERSON, Ph.D. Washington 1984. Optimal choice of regulatory instrument in a fishery under uncertainty and instrument adjustment constraints.

RAGNAR ARNASON, Ph.D. British Columbia 1984. Efficient harvesting of fish stocks: The case of the Icelandic demersal fisheries.

CYNTHIA S. BANTILAN, Ph.D. North Carolina State 1984. A disequilibrium analysis of the market for agricultural credit in low-income countries.

JUAN C. BATISTA, Ph.D. Cornell 1985. Regional opportunity costs of producing milk and the agricultural production effects of alternative spatial pricing policies in dairying.

RUTH BEN-ZVI, Ph.D. SUNY-Stony Brook 1985. Optimal production of oil.

TROND BJØRNDAL, Ph.D. British Columbia 1985. The optimal management of an ocean fishery.

DARRELL BOSCH, Ph.D. Minnesota 1984. The value of soil, water, and weather information in increasing irrigation efficiency.

CAROL E. BRAY, Ph.D. Stanford 1984. A quantitative analysis of price integration in the cotton market by quality and region.

GERALDO M. CALEGAR, Ph.D. Minnesota 1984. Brazilian wheat policy and its income distribution and trade effects: A case study.

LAURIE A. CAMERON, Ph.D. Oregon 1985. The commercial demand for electricity in the Pacific Northwest: A disaggregated, end use approach.

ROBERT N. COLLENDER, Ph.D. California-Berkeley 1984. Optimal land allocation under uncertainty: A generalized portfolio approach.

BEVERLY A. COOK, Ph.D. Simon Fraser 1984. Optimum harvest levels for Canada's Pacific halibut fishery.

PAUL A. COOMES, Ph.D. Texas-Austin 1985. Stochastic control and U.S. agricultural policy.

LALITH DESILVA, Ph.D. George Washington 1985. Residential fuelwood demand analysis in an *LDC* context: A case study of Sri Lanka.

PRAVEEN M. DIXIT, Ph.D. Purdue 1984. Endogenous policy formation: An econometric and simulation analysis of the U.S. coarse grains sector in an open economy.

DEIVAN DURAI, Ph.D. Kansas State 1985. Analysis of returns to agricultural research on wheat and wheat productivity in the Great Plains, 1966 to 1980.

KAREN A. DVORÁK, Ph.D. California-Berkeley 1984. Soil productivity and crop production.

EDWARD H. EASTERLING, Ph.D. Missouri-Columbia 1984. Locational factors affecting the broiler industry.

CYPRIAN EJIASA, Ph.D. Ohio State 1985. The exchange rate and the competitiveness of U.S. agricultural commodity trade.

MICHAEL D. FRANK, Ph.D. Illinois-Urbana 1984. Structural change in the retail demand for beef, chicken, and pork: An application of gradual switching regression.

EDWARD O. FRYAR, JR., Ph.D. Minnesota 1983. Edible vegetable oil prices in the U.S. and world markets.

HELEN GAGLIA, Ph.D. Queen's 1985. Environmental quality and optimal allocation of resources in a dynamic two-sector economy.

VILLAMOR B. GAMPONIA, Ph.D. Washington 1984. Optimal rotation and search for a growing asset.

VICTOR GANOZA, Ph.D. North Carolina State 1985. Estimating the value of fixed assets used in North Carolina dairy enterprises.

WAYNE GINEO, Ph.D. Minnesota 1984. Agricultural income insurance: A policy alternative.

GLADYS M. GONZALEZ, Ph.D. Missouri-Columbia 1984. Conversion of corn belt agricultural land to irreversible uses: Measuring increased production costs.

KAREN GULLIVER, Ph.D. Minnesota 1984. The Brazilian soybean economy: An econometric model with emphasis on government policy.

YOSHINOBU HABUKI, Ph.D. Washington State 1984. The U.S. hop industry: Structural analysis and forecasts.

STEPHEN L. HALEY, Ph.D. Purdue 1985. The theory of agricultural comparative advantage.

JOHN M. HALLORAN, Ph.D. Michigan State 1984. An economic evaluation of computerized markets in future delivery contracts as a coordination mechanism for the U.S. potato subsector.

SERMIN D. HARDESTY, Ph.D. California-Davis 1984. Impact of the 1981 tax act: A dynamic analysis of farm firm production, investment, and financing decisions.

KIM S. HARRIS, Ph.D. Illinois-Urbana 1985. A projective method for measuring central Illinois farmers' reservation prices for cash and intermediate credit reserves.

MARY G. HASELTON, Ph.D. Pennsylvania State 1985. An econometric analysis of industrial fuel use in the 1970's.

STEPHEN A. HATCHETT, Ph.D. California-Davis 1984. Dynamic input decisions: An econometric analysis of crop response to irrigation.

PETER J. HEFFERNAN, Ph.D. Ohio State 1984. An analysis of interest rate hedging strategies of agricultural lenders.

DANA L. HOAG, Ph.D. Washington State 1984. An evaluation of USDA commodity program incentives for erodible land retirement.

JACK E. HOUSTON, JR., Ph.D. Washington State 1984. Water and energy conservation modeling in Pacific Northwest irrigated agriculture.

MAHMOOD IQBAL, Ph.D. Simon Fraser 1985. An analysis of energy in Pakistan, 1960–81.

DAVID IREFIN, Ph.D. Washington State 1984. An application of mixed integer programming to selection of least-cost machinery complements for the Columbia Basin Project area farms.

WILLIAM K. JAEGER, Ph.D. Stanford 1985. Agricultural mechanization: The economics of animal traction in Burkina Faso.

DUSIT JESDAPIPAT, Ph.D. Illinois-Urbana 1984. Policy and structures of Japanese import demand for beef and feedgrains.

MUTAWAKKIL KAZI, Ph.D. Boston 1985. Determinants of agricultural porductivity in the lower Indus Basin: An empirical analysis of agriculture in Sind Province.

ROBERT C. KEEN, Ph.D. Purdue 1985. Comparative performance of cooperative and proprietary grain and farm supply firms.

ROBERT KELLOGG, Ph.D. North Carolina State 1985. A bioeconomic model for determining the optimal timing of harvest with application to two North Carolina fisheries.

MOKIWA A. KIGODA, Ph.D. Missouri-Columbia 1984. Incentives for food crop production in Tanzania: With special reference to the Mbeya region.

THOMAS O. KNIGHT, Ph.D. Missouri-Columbia 1984. Evaluation for extension programs using subjective probability distributions and normative decision models.

MARY F. KOKOSKI, Ph.D. North Carolina-Chapel Hill 1984. A general equilibrium analysis of the measurement of the economic impacts of climatic changes.

DANIEL E. KUGLER, Ph.D. Michigan State 1984. Variable-cost-sharing level program implications for Kentucky's Jackson Purchase area: An economic and policy study of cash grain production considering soil depletion.

EUGENE L. KUNDA, Ph.D. Illinois-Urbana 1984. Analysis of the transmission of market information among a sample of Illinois grain elevators.

ERIK R. LICHTENBERG, Ph.D. California-Berkeley 1985. The role of land quality in agricultural diversification.

KUO-CHING LIN, Ph.D. North Carolina State 1984. The effect of expected inflation on real land values.

CHARLES R. LINDSEY, Ph.D. Princeton 1984. Import disruptions, exhaustible resources, and intertemporal security of supply.

MANUEL LOPEZ-BLANCO, Ph.D. Michigan State 1984. A multipurpose simulation model of dairy/forage farming systems in northwestern Spain.

DONALD R. MCDOWELL, Ph.D. Illinois-Urbana 1985. Conventional production functions vs. energy term production functions: Energy used for crop production in Midwest agriculture.

STEPHEN D. MCGARY, Ph.D. Washington State 1984. An intraseasonal price forecasting model for Washington fresh apples.

MICHELE C. MARRA, Ph.D. North Carolina State 1984. Examining double cropping with utility theory.

LEWIS C. MAY, Ph.D. Washington State 1984. The cost-size relationship revisited: Dryland small grain farms in the Pacific Northwest.

GEORGE J. MERGOS, Ph.D. Stanford 1985. Price, technical change, and output composition effects on derived demand for grains: A cost function approach.

JEFFREY C. METZEL, Ph.D. Fletcher School 1985. The integration of irrigated agriculture into traditional farm systems in the upper Senegal River valley: A linear programming analysis of economic variable in farm production.

FAIZ MOHAMMAD, Ph.D. Simon Fraser 1984. An analysis of the structure and performance of agricultural markets in Pakistan.

BRIAN J. MORTON, Ph.D. California-Berkeley 1985. Coal leasing in the fourth world: Hopi and Navajo coal leasing, 1954–77.

ESHYA MUJAHID, Ph.D. Boston 1985. Effects of policies on agricultural development: A case study of the Punjabs and Bengals.

SAEED NABAEE-TABRIZ, Ph.D. Ohio State 1985. An economic analysis of soil conservation limitations on the intensity of cropland use in Ohio.

ABBAS NAJAFIZADEH, Ph.D. California-Santa Barbara 1985. The structure of the world oil market and the role of OPEC.

PETER H. NICKERSON, Ph.D. Washington 1984. The effects of regulation on lottery participation: The case of big game hunting in Washington state.

CHARLES M. OELLERMANN, Ph.D. Purdue 1985. Futures and cash price relationships for beef cattle.

ABDALLAH OMEZZINE, Ph.D. Missouri-Columbia 1985. Market ananlysis of the winter vegetable industry in Tunisia.

DAVID R. ORDEN, Ph.D. Minnesota 1984. Capital flows, the exchange rate, and agricultural commodity markets.

SYLVESTER I.I. OREWA, Ph.D. Kansas State 1984. Labor shortage in Nigerian plantations: An economic evaluation of early years intercropping and tree replacement policy in oil palm plantations.

CLAUDIA A. PARLIMENT, Ph.D. California-Berkeley 1984. Agricultural production cooperatives: Factors affecting performance.

ROBERT H. PATRICK, Ph.D. New Mexico 1985. Peak-load pricing with continuous interdependent demand.

JOANN E. PAULSON, Ph.D. Stanford 1984. The structure and development of rural financial markets in Kenya: An example of agricultural finance.

DAVID A. PENN, Ph.D. Oklahoma 1984. An analysis of the economic and ideological factors involved in the development of rights to water in the West, focusing on Colorado.

NGUYEN V. QUYEN, Ph.D. Western Ontario 1984. The exploitation of a nonrenewable resource under imperfect competition.

JERONIMO RAMOS, Ph.D. Minnesota 1984. An economic evaluation of government intervention in the Mexican agriculture sector: The corn and wheat sectors.

ROBERTO DE LA FUENTE RANOLA, JR., Ph.D. Minnesota 1984. Cropping systems and determinants: Factors influencing cropping patterns, fertilizer-use and resource-use efficiency.

THOMAS A. REARDON, Ph.D. California-Berkeley 1984. Agricultural price policy in Peru.

PAUL C. RIETHMULLER, Ph.D. Minnesota 1984. Government intervention in commodity markets: The case of the Japanese rice and wheat markets.

MARY A. ROSS, Ph.D. Illinois-Urbana 1985. International trade in selected processed food and beverages.

RANDAL R. RUCKER, Ph.D. Washington 1984. An economic analysis of bidding and cutting behavior on public timber sales contracts.

FREDERICK J. RUPPEL, Ph.D. Maryland 1984. Export sales and export shipments of corn, wheat, and soybean: Institutional relationships and empirical evidence.

LINDA M. SALCHENBERGER, Ph.D. Northwestern 1985. A two-stage differential game of renewable resources.

LARRY D. SANDERS, Ph.D. Colorado State 1985. economic benefits of river protection: A study of recreation use and preservation values.

FENTON B. SANDS, Ph.D. Michigan State 1984. An economic analysis of farm-level livestock marketing in eastern Upper Volta.

CARLOS A. M. SANTANA, Ph.D. Minnesota 1984. The impact of economic policies on the soybean sector of Brazil: An effective protection analysis.

EMMANUEL S. SANTIAGO, Ph.D. Kansas State 1984. A comparative analysis of costs and effectiveness of simulated food security programs for the association of southeast Asian nations.

CLAUDIO SAPELLI, Ph.D. Chicago 1985. Government policy and the Uruguayan beef sector.

MUSTAFA A. SAWANI, Ph.D. Missouri-Columbia 1985. Some effects of capital rationing and interest rates on farm organization and income.

RICHARD J. SEXTON, Ph.D. Minnesota 1984. The formation of cooperatives: An analysis of entry incentives, entry deterrence, optimal financing arrangements, and stability requirements.

TERENCE C. SHEALES, Ph.D. Cornell 1985. Hedging

Australian wheat exports on U.S. futures markets: An analysis of potential gains and losses.

MICHAEL SHELBY, Ph.D. Boston 1985. Analysis of recent air quality policies in the United States to limit sulfur dioxide emissions to the environment.

JACQUELINE R. SHERMAN, Ph.D. Michigan 1984. Grain markets and the marketing behavior of farmers: A case study of Mang, Upper Volta.

WILLIAM W. SMITH, Ph.D. Cornell 1985. The effects of reforming electricity rates on employment levels in New York State: A multisectoral analysis.

UKAR W. SOLESTIJO, Ph.D. West Virginia 1984. Evaluation of the potential economic benefit of coal-oil substitution in the Indonesian economy: An interindustry approach.

DALE E. SQUIRES, Ph.D. Cornell 1984. Resource economics regulation and the structure of production in a multispecies fishery: A study of the New England otter trawl industry.

RIGOBERTO STEWART, Ph.D. North Carolina State 1984. Basic grains pricing policies and their effects in Costa Rica.

WILLIAM J. STRANG, Ph.D. Washington 1984. Models of environmental externalities.

CHONG H. SUH, Ph.D. Missouri-Columbia 1984. Factors influencing off-farm employment and land-use intensity in Korea: A simultaneous equations approach.

HABIBAH SULEIMAN, Ph.D. Missouri-Columbia 1985. An econometric model of the Malaysian natural rubber industry: A phenological approach.

JOSEPH R. SZMANIA, Ph.D. New York 1985. A multifuel analysis of energy consumption and appliance choice in the residential sector.

THOMAS P. TOMICH, Ph.D. Stanford 1984. Private land reclamation in Egypt: Studies of feasibility and adaptive behavior.

DAVID D. TRECHTER, Ph.D. Michigan State 1984. An economic analysis of farm revenue insurance.

ROBERTO D. VELLUTINI, Ph.D. Cornell 1985. An economic analysis of agricultural policy tradeoffs in São Paulo, Brazil.

WOLFGANG O. VOGEL, Ph.D. Minnesota 1984. An economic analysis of the potential sunflower acreage in the upper Midwest.

FREDERICK W. H. WACKERNAGEL, Ph.D. Cornell 1985. Rice for the terraces: Cold-tolerant varieties and other strategies for increasing rice production in the mountains of southeast Asia.

MELLIE L. WARNER, Ph.D. North Carolina State 1984. A portfolio model of wheat acreage.

NANCY A. WILLIAMS, Ph.D. California-Berkeley 1984. Iterative planning with incentive compatible control: The case of the USDA Forest Service.

FREDERICK D. WORMAN, Ph.D. Kansas State 1985. Utilizing agronomic crop growth models in economic analysis: The case of cropping adjustments to decreasing irrigation water availability in western Kansas.

DAVID WING-KAY YEUNG, Ph.D. York 1985. Optimal exploitation of nonrenewable resources in stochastic environments.

THOMAS P. ZACHARIAS, Ph.D. Illinois-Urbana 1984. Economically optimal integrated pest-management strategies for control of corn rootworm and soybean cyst nematode.

GEOFFREY L. ZIMMERMAN, Ph.D. Colorado-Boulder 1985. The capitalization of climatological and topographical amenities in the regional earnings of recent U.S. migrants.

Manpower, Labor, and Population, including Trade Unions and Collective Bargaining

ISSA H. AL-QUISI, Ph.D. Washington 1984. Earning determination and differential in the Kuwait market.

ESKANDER ALVI, Ph.D. Johns Hopkins 1985. Three essays on the theory of unemployment.

BETH J. ASCH, Ph.D. Chicago 1984. Pension, wages, and sorting: A theoretical and empirical analysis.

JOHN G. BAUER, Ph.D. Michigan 1984. Age at marriage and earnings in Peninsular Malaysia.

PAUL BEAUMONT, Ph.D. Pennsylvania 1984. ECESIS: An interregional economic-demographic model of the United States.

WILLIAM K. BELLINGER, Ph.D. Northwestern 1985. Studies in collective bargaining.

JEFF E. BIDDLE, Ph.D. Duke 1985. Empirical investigations of life cycle labor supply models.

EMILY C. BLANK, Ph.D. Boston College 1984. The effect of layoffs and quits on wage growth of male household heads.

WILLIAM M. BOAL, Ph.D. Stanford 1985. Unionism and productivity in West Virginia coal mining.

MARY O. BORG, Ph.D. North Carolina-Chapel Hill 1985. An economic model of fertility in Korea.

HOWARD BOTWINICK, Ph.D. New School 1985. Wage differentials and the competition of capitals: A new explanation for inter- and intra-industry wage differentials among workers with similar skill levels.

LINDA CAVALLUZZO, Ph.D. SUNY-Buffalo, 1984. An evaluation of the nonpecuniary returns to labor: Navy pilots as a test case.

CHIRASAK CHUNTAPANICH, Ph.D. City-New York 1985. Determinants of labor supply in Thailand.

THOMAS S. COLEMAN, Ph.D. Chicago 1984. Essays on aggregate labor market business cycles fluctuations.

JOHN L. CONANT, Ph.D. Tennessee-Knoxville 1984. The economics of labor union mergers.

ALEJANDRA COX EDWARDS, Ph.D. Chicago 1984. Three essays on labor markets in developing countries.

SUSAN DAVIS, Ph.D. New School 1985. The impact of economic growth on women: A case study of Texas, 1960–80.

JAYNE DEAN, Ph.D. New School 1985. A dual wage structure approach to occupational and industrial segregation and wage differentials by sex.

TODD E. EASTON, Ph.D. California-Berkeley 1984. The impact of collective bargaining on teacher salary determination in Oregon public schools.

LISA EHRLICH, Ph.D. Pennsylvania 1984. Women's career orientation, labor supply, and fertility behavior.

STANLEY W. ELSEA, D.B.A. Indiana-Bloomington 1984. Controlling absenteeism by means of policy variables: An empirical investigation.

WILLIAM EVEN, Ph.D. Iowa 1984. The relationship between married women's labor supply and fertility decisions.

ROBERT A. FELDMAN, Ph.D. Princeton 1985. An examination of health status measurement in empirical models of retirement behavior.

JUDITH FIELDS, Ph.D. New York 1984. Industry mix, the demand for skill, and women's earnings.

THOMAS P. FRAZIER, Ph.D. American 1984. Consultants, unions, and NLRB elections.

MARION T. GHILARDUCCI, Ph.D. California-Berkeley 1984. Pensions and collective bargaining: Towards a national policy on retirement income support.

PAUL W. GLEWWE, Ph.D. Stanford 1985. An analysis of income distribution and labor in Sri Lanka.

STEPHAN F. GOHMANN, Ph.D. North Carolina State 1984. Social Security benefit acceptance and labor supply.

JANET HOLTZBLATT, Ph.D. Wisconsin-Madison 1984. The effects of plant closings on workers' earnings and transfer receipts.

AVERY HOROWITZ, Ph.D. City-New York 1984. Determinants of fertility in the United States: An analysis based on the 1980 Census of Population.

DAVID R. HOWELL, Ph.D. New School 1985. Production technology, job skills, and relative industry wages: A study of production worker wages in U.S. manufacturing industries.

WEI-CHIAO HUANG, Ph.D. California-Santa Barbara 1984. A study of the indirect immigration of professional manpower to the United States.

LAURENCE R. IANNACCONE, Ph.D. Chicago 1984. Consumption capital and habit formation with an application to religious participation.

JOHN M. JEFFERIES, Ph.D. New School 1985. Factors influencing minority choices of occupation-specific training at the postsecondary school level.

CAROLINE LI-MING JEN, Ph.D. SUNY-Buffalo 1984. Urban female labor force participation in Brazil.

THOMAS M. KARIER, Ph.D. California-Berkeley 1985. The union impact on profits, productivity, and prices.

DILSHAD KHAWAJA, Ph.D. City-New York 1985. Determinants of fertility and child quality in Pakistan.

JAE WON KIM, Ph.D. Indiana-Bloomington 1984. Estimation of the earnings functions of married women in the presence of discontinuous labor supply.

JANG-HO KIM, Ph.D. Northwestern 1985. Firm-specific human capital, bonus payments, and interscale earnings differentials in Korea.

RICHARD KOLSKY, Ph.D. Yale 1985. The youth employment puzzle: Putting together the pieces.

JOUNG-YONG LEE, Ph.D. Maryland 1985. The baby boom generation: Its impact on relative earnings of young workers and earnings inequality among all workers.

ANAT LEVY, Ph.D. California-Los Angeles 1984. Numbers and sizes of unions.

MARIAN E. LEWIN, Ph.D. Wisconsin-Madison 1985. A microeconomic analysis of wage-indexed contracts.

ANTON D. LOWENBERG, Ph.D. Simon Fraser 1985. Towards an economic theory of the apartheid state.

FEI-YING LU, Ph.D. Northwestern 1985. The dynamics of family formation and population duality.

SAEID MAHDAVI, Ph.D. California-Santa Barbara 1985. Population and socioeconomic development: An empirical analysis of cross-national differentials in fertility, female labor force participation, and size distribution of income.

STEPHEN L. MANGUM, Ph.D. George Washington 1984. Issues in the economics of training with application to the U.S. Army.

DONALD W. MAR, Ph.D. California-Berkeley 1985. Chinese imigrants and the ethnic labor market.

FELICE F. MARTINELLO, Ph.D. British Columbia 1984. Wage and employment determination in a unionized industry: The IWA in the B.C.. wood products industry.

JANE E. MATHER, Ph.D. Stanford 1984. A study of cyclical wage flexibility using disaggregated data.

ROBERT N. MAYER, Ph.D. Northwestern 1985. The human resources function as a labor market intermediary.

ALLECHI M'BET, Ph.D. Wisconsin-Milwaukee 1985. Minimum wage, inflation, and unemployment: A simultaneous equations analysis.

KURT MIKAN, Ph.D. Alabama 1985. An impact evaluation of the training of Oak Ridge associated universities and the Nuclear Division of Union Carbide Corporation.

MICHAEL J. MOORE, Ph.D. Michigan 1984. Three essays in labor economics: Models of labor market equilibrium.

THOMAS A. MROZ, Ph.D. Stanford 1984. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions.

MACHIKO OSAWA, Ph.D. Southern Illinois-Carbondale 1984. Human capital formation, labor force participation, and fertility of Japanese women.

JAMES H. PEOPLES, Ph.D. California-Berkeley 1984. The expected impact of deregulation on labor earnings: The long distance telecommunication industry.

MONTGOMERY PEREIRA Ph.D. North Carolina State

1985. Economic analysis of land and labor market participation in rural India: An application of the multinomial logit model.

WENDY L. RAYACK, Ph.D. Wisconsin-Madison 1984. The distributional impacts of recessions: An analysis of earnings sensitivity by long-run family income class.

WILLIAM R. REED, Ph.D. Northwestern 1985. An analysis of nonpecuniary job attributes as determinants of workers' quit behavior.

MARSHALL B. REINSDORF, Ph.D. Maryland 1985. Retirement, partial retirement, and minimum hours constraints.

JAMES C. ROBINSON, Ph.D. California-Berkeley 1984. Work and health: An economic and policy analysis.

RICHARD D. ROGERSON, Ph.D. Minnesota 1984. Topics in the theory of labor markets.

RENE ROSENBAUM, Ph.D. Notre Dame 1985. Tripartite collective bargaining: The case of farmers, farmworkers, and processors in the Ohio processing tomato industry.

CHRISTOPHER J. RUHM, Ph.D. California-Berkeley 1984. Job tenure and the consequences of job separations.

JOHN F. SCHNELL, Ph.D. Illinois-Urbana 1984. The regulation of work: Seniority rules and constrained choice.

ROBERT H. SCHWARTZ, Ph.D. Michigan 1983. An exploration of Holland's theory for use in the selection of dental students.

ABDULLA I. SHAMIA, Ph.D. Washington 1984. Forecasting the Ph.D. labor market: Econometric analysis.

PETER D. SHERER, Ph.D. Wisconsin-Madison 1985. Determinants of salary raise decisions.

VILIS H. SKULMIS, Ph.D. Queen's 1984. Unemployment spell durations: A closer look.

AHMAD SOHRABIAN, Ph.D. California-Santa Barbara 1984. Intertemporal substitution hypothesis and labor market fluctuations.

SALLY J. W. SOMMER, Ph.D. Colorado-Boulder 1985. An economic analysis of the migratory farm labor market in northwest Ohio.

LEE SPECTOR, Ph.D. Iowa 1985. A macroeconomic examination of changes in the size and composition of the labor force.

STEVEN N. STERN, Ph.D. Yale 1985. Search, applications, vacancies, and equilibrium labor markets.

BARRY M. STREGEVSKY, Ph.D. Illinois-Urbana 1985. A comparative study of labor force participation of native and immigrant married women.

J. EDWARD TAYLOR, Ph.D. California-Berkeley 1984. Migration networks and risk in household labor decisions: A study of migration from two Mexican villages.

NASSER VASEGH-DANESHVARY, Ph.D. Tennessee-Knoxville 1984. Economic determinants of immigration, initial location, and remigration of foreign-born trained manpower in the United States.

RAFAEL E. VILLAGOMEZ, Ph.D. Oklahoma State 1985. Migration and quality of life: A study at the SMSA level.

STEPHEN VOSTI, Ph.D. Pennsylvania 1984. Nutrition, health, and wages in rural south India.

YUNG-CHANG WANG, Ph.D. SUNY-Stony Brook 1985. An economic analysis of fertility behavior in Taiwan.

DOUGLAS R. WHOLEY, Ph.D. California-Berkeley 1984. Organizational influences on career patterns: An analysis of law firms.

DONALD R. WILLIAMS, Ph.D. Northwestern 1984. Racial differences in the labor force participation of male teenagers.

EVA CHUNG-CHIUNG YEN, Ph.D. Washington-St. Louis 1985. Determinants of fertility in the LDCs: The household production approach revisited.

Welfare Programs; Consumer Economics; Urban and Regional Economics

JOSEPH BAIGES PLANAS, Ph.D. California-Los Angeles 1984. General equilibrium with local public goods: The role of land, property taxation, and zoning laws.

KAPIL BAWA, Ph.D. Columbia 1984. Modeling inertia and variety-seeking tendencies in brand choice behavior.

DENNIS BECKMANN, Ph.D. Iowa 1984. The effect of transportation cost increases on residential density and employment opportunities.

EMIL B. BERENDT, Ph.D. City-New York 1985. The demand for AFDC and Medicaid.

DAVID E. BERRY, Ph.D. Wisconsin-Madison 1984. The effects of youth employment opportunities and public school per pupil expenditure on high school and college enrollment.

DIXIE BLAKELY, Ph.D. Syracuse 1984. Housing demand and residential location in the Chicago SMSA: One- and two-worker households in a multicentric model.

PETER G. BOHMER, Ph.D. Massachusetts-Amherst 1985. The impact of public sector employment on racial inequality, 1950-84.

DAVID BURRESS, Ph.D. Wisconsin-Madison 1985. The interest elasticity of household saving under selected models of intertemporal optimization: A microeconomic method.

JAE H. BYON, Ph.D. Washington 1984. Moral hazard and the health maintenance organization.

WENDY A. CAMPIONE, Ph.D. Illinois-Urbana 1984. Land conversion at the urban fringe: An investigation into effects of differential property tax assesment.

YUVAL COHEN, Ph.D. Columbia 1985. Nonuniform values of time and congestion toll pricing in bottleneck and queue situations.

RICHARD L. COOPERSTEIN, Ph.D. Maryland 1985. Quantifying the decision to become a first-time home buyer.

GARY M. CRAKES, Ph.D. Connecticut 1984. An econom-

ic estimation of the dental practice production process.

ROBERT CRUZ, Ph.D. Pennsylvania 1985. Macro modeling of the urban economy: Theoretical considerations and an empirical application to south Florida.

WILLIAM J. FURLONG, Ph.D. Queen's 1984. Crime and punishment: A general equilibrium analysis.

YITCHAK HABERFELD, Ph.D. Wisconsin-Madison 1984. Implications of measurement error in human capital studies of gender-based wage discrimination.

DAVID E. HARRINGTON, Ph.D. Wisconsin-Madison 1984. An intertemporal model of housing demand.

ADELE M. HAYUTIN, Ph.D. California-Berkeley 1984. Scale economics in highway capacity: Empirical evidence and policy implications.

ROBERT W. HELSLEY, Ph.D. Princeton 1985. Economies of agglomeration and urban spatial structure.

SUSAN HOLAK, Ph.D. Columbia 1985. An approach for analyzing innovative consumer durables: An intermediate test procedure for suggesting product modification and predicting acceptance.

JUDITH C. HUSHBECK, Ph.D. American 1984. An economic history of old age in the United States: Occupational obsolescence and downward mobility, 1860–1920.

THEODORE JOYCE, Ph.D. City-New York 1985. A structural model of birth outcomes in the United States.

HENDRIKUS KEMNA, Ph.D. Pittsburgh 1985. The relationship between schooling and health: Investigation into an occupational linkage.

AMIR KIA, Ph.D. Carleton 1985. Evaluating regional policies in Canada: The case of Newfoundland.

SOONWON KWON, Ph.D. Johns Hopkins 1985. Demand for medical care among the rural poor in Korea: An empirical investigation using survey data.

JOANNE S. LAVIN, Ph.D. Columbia 1984. Geographic aggregation in physical distribution problems.

JOHN C. LEADLEY, Ph.D. Wisconsin-Madison 1985. Estimating the costs and benefits of the 1962 FDA amendments.

KYUBANG LEE, Ph.D. Ohio State 1985. The demand for owner-occupied housing: A study of the simultaneity among housing demand, the choice of loan-value ratio, and the length of stay.

JOHN R. LOTT, JR., Ph.D. California-Los Angeles 1984. Alternative explanations for public provision of education.

CHARLES H. MARSTON, Ph.D. California-Berkeley 1985. Urban efficiency advantages.

JEFFREY L. MEMMOT, Ph.D. Texas A&M 1984. Economic evaluation model of highway improvement projects.

HUSSEIN MIRGHANI, Ph.D. Syracuse 1985. Costs and benefits of malaria control in the Gezire scheme of the Sudan.

SAM MIRMIRANI, Ph.D. Clark 1985. The impact of certificate-of-need and Section 1122 of Social Security Act on investment of hospital industry—long run.

HARRIET NEWBURGER, Ph.D. Wisconsin-Madison 1984. Discrimination by real estate brokers in response to white prejudice.

GEORGE R. PARSONS, Ph.D. Wisconsin-Madison 1984. Estimating demand functions for attributes of a differentiated product.

MICHAEL RADZICKI, Ph.D. Notre Dame 1985. The impact of an increased level of vocational and technical education on employment and population growth in the Indiana economy.

TERESA RILEY, Ph.D. Syracuse 1984. The impact of the antidiscrimination legislation on the employment status of women.

JOHN RIZZO, Ph.D. Brown. Health care policy and heterogeneous risk group: Theory and a U.S.–Japan comparison.

CLAUDIA L. SCHUR, Ph.D. Maryland 1985. Imperfect information in health insurance markets: Reconciling theory and observation.

VIBHOOTI SHUKLA, Ph.D. Princeton 1984. The productivity of Indian cities and some implications for development policy.

ANNE C. SKELLEY, Ph.D. Brown. Essays on rent control.

LYNN SMITH, Ph.D. Pittsburgh 1985. Regional differences in the level of unionization in the United States: A study of the effects of right-to-work laws in a dynamic setting.

RICHARD N. SPIVACK, Ph.D. Connecticut 1984. The determinants of housing abandonment: A case study of Providence, Rhode Island.

K. L. STEIN, Ph.D. Temple 1985. Age discrimination in employment: An economic inquiry into the employer's decision to discriminate—A case study.

RICHARD A. STOUT, Ph.D. Indiana-Bloomington 1984. Optimal landlord behavior under rent controls.

THOMAS TAYLOR, Ph.D. Iowa 1984. A structural model of the market for medical care services.

RENEE D. TOBACK, Ph.D. Massachusetts-Amherst 1985. Protective labor legislation for women: The Massachusetts ten-hour law.

D. T. WARGO, Ph.D. Temple 1985. A regional disaggregrated econometric model of the housing market, mortgage market and deposit market.

VIRGINIA L. WILCOX-GOK, Ph.D. Washington-St. Louis 1984. Sibling data and family influences on child health.

ALLEN J. WILKINS, Ph.D. Wisconsin-Madison 1984. The economics of heroin addiction and criminal activity.

JAMES T. WILKINSON, Ph.D. Vanderbilt 1985. Alcohol and accidents: An economic approach to drunken driving.

JONG G. YOO, Ph.D. Northern Illinois 1985. An empirical investigation of the individual welfare inequality in Korea: 1965–83.

THE AMERICAN ECONOMIC REVIEW

VOLUME LXXV

THE AMERICAN ECONOMIC ASSOCIATION

Executive Office: Nashville, Tennessee

Editorial Office: 169 Nassau Street, Princeton, NJ 08542-7067

CONTENTS OF ARTICLES AND SHORTER PAPERS

CONTENTS OF THE PAPERS AND PROCEEDINGS

CONTRIBUTORS TO ARTICLES AND SHORTER PAPERS

CONTRIBUTORS TO PAPERS AND PROCEEDINGS

Computer Access to Articles in the JEL Subject Index

On-line computer access to the *JEL* and *Index of Economic Articles* data base of journal articles is currently available through DIALOG Information Retrieval Service. DIALOG file 139 *(Economic Literature Index)* contains complete bibliographic citations to articles from some 260 journals listed in the quarterly *JEL* issues from 1969 through the current issue. The abstracts in *JEL* since June 1984 are also available as part of the full bibliographic record. The file may be searched using free-text searching techniques or author, journal, title, geographic area, date, and other descriptors, including descriptor codes based on the *Index's* four-digit subject classification numbers.

DIALOG offers a variety of contract choices, including the option to pay for only what you use—*no minimum, no initiation or start-up fee.* Most university libraries already subscribe to DIALOG. For information on the DIALOG service, contact your librarian or write to or call:

DIALOG Information Services, Inc., Marketing Department, 3460 Hillview Avenue, Palo Alto, California 94304 (800/227-1927 or 800/982-5838, in California, or 415/858-3785).

The American Economic Review

ARTICLES

SEPTEMBER 1985

THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

Officers

President
CHARLES P. KINDLEBERGER
Massachusetts Institute of Technology

President-Elect
ALICE M. RIVLIN
The Brookings Institution

Vice Presidents
ELIZABETH E. BAILEY
Carnegie-Mellon University
JOSEPH E. STIGLITZ
Princeton University

Secretary
C. ELTON HINSHAW
Vanderbilt University

Treasurer
RENDIGS FELS
Vanderbilt University

Managing Editor of The American Economic Review
ORLEY ASHENFELTER
Princeton University

Managing Editor of The Journal of Economic Literature
MOSES ABRAMOVITZ
Stanford University

Executive Committee

Elected Members of the Executive Committee
WILLIAM D. NORDHAUS
Yale University
A. MICHAEL SPENCE
Harvard University
VICTOR R. FUCHS
Stanford University
JANET L. NORWOOD
Bureau of Labor Statistics
ALAN S. BLINDER
Princeton University
DANIEL L. McFADDEN
Massachusetts Institute of Technology

EX OFFICIO Members
W. ARTHUR LEWIS
Princeton University
CHARLES L. SCHULTZE
The Brookings Institution

•Printed at Banta Company, Menasha, Wisconsin. The publication number is ISSN 0002-8282.

•*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the Directory, and Supplements, is published by the American Economic Association and is sent to all members six times a year: March; May; June; September; semi-monthly, December.

Regular member dues for 1985, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature* are as follows:

$35.00 if annual income is $30,000 or less;
$42.00 if annual income is above $30,000, but no more than $40,000;
$49.00 if annual income is above $40,000.

Nonmember subscriptions will be accepted only for both journals: Institutions (libraries, firms, etc.), $100 a year; individuals, $65.00. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add $11.00 to the annual rates above to cover extra postage.

•Correspondence relating to the Directory, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

•Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

•Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

ALBERT O. HIRSCHMAN

DISTINGUISHED FELLOW

1984

The thread that runs through the work of Albert Hirschman is the effect of imbalance on growth. This interest is already present in his first book, *National Power and the Structure of Foreign Trade* (1945), which explores the effects of unequal power in foreign trade. Subsequent books deal mainly with domestic markets in less developed countries, where Hirschman focuses on the stimulus that comes from shortages, bottlenecks, and other stressful mechanisms. In such situations, linkages call forth resources and abilities that are hidden, scattered, or badly utilized. *The Strategy of Economic Development* (1958) centers on the shortage of entrepreneurial experience, and also on the latitude that different operations allow for poor performance. Later writing concentrates on the pressures created by consumers, in choosing between *Exit, Voice and Loyalty* (1970). This leads Hirschman into a theory of consumer goods and services turning not on prices but on the tendency of the thing consumed to cause disapppointment rather than satisfaction (*Shifting Involvements*, 1982).

Hirschman's immense knowledge of the history of social ideas illuminates all his writing, and is the basis for his brilliant monograph, *The Passions and the Interests* (1977). This too incorporates his interest in the "unintended effects of human action," but deals mainly with the transformation of the pursuit of economic interests from being a sinful to being a virtuous way of life. The eminence of Hirschman is testified by the great flow of writing in several disciplines in response to his ideas.

THE AMERICAN ECONOMIC REVIEW

• Submit manuscripts (4 copies), no longer than 50 pages, double spaced, to:
Orley Ashenfelter, Managing Editor, *AER*; 169 Nassau Street, Princeton, NJ 08542–7067.

• Submission fee: $25 for members; $50 for nonmembers. Style guides will be provided upon request.

• No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

September 1985

VOLUME 75, NUMBER 4

Articles

Shorter Papers

Animals' Choices over Uncertain Outcomes: Some Initial Experimental Results

By RAYMOND C. BATTALIO, JOHN H. KAGEL, AND DON N. MACDONALD*

This paper reports experiments dealing with rats' choices between uncertain prospects. The experiments take as their starting point that in choosing between prospects with discriminable outcomes, one of which dominates the other according to a first-degree stochastic dominance criteria, rats prefer the dominant alternative. Evidence on this point is twofold. First, in choosing between two alternatives which are identical with respect to quantity and quality of reinforcement, but which differ with respect to probability of reinforcement, rats clearly prefer (choose more frequently) the alternative with the higher probability of reinforcement. Preferences become more extreme as differences in the probability of reinforcement increase, rapidly approaching nearly exclusive preference for the alternative with the higher probability of payoff (see Norman Sutherland and Nicholas Mackintosh, 1971). This is a robust result found in numerous experiments with different species of animals.[1] Second, in choosing between two outcomes which differ with respect to quantity but not quality or probability of reinforcement, rats (and other animals) prefer the dominant alternative. This result is sufficiently robust that virtually all behavioral models predict it, whether they incorporate maximizing principles or take pains to separate themselves from maximizing formulations. Studies here have been limited almost exclusively to certain prospects, undoubtedly because the extension to uncertain outcomes appears too trivial to pursue.

With the notable exception of some recent research by Thomas Caraco (1981, 1982) and Caraco et al. (1980), only a handful of papers have been reported which extend the study of animals' choices over uncertain prospects. These have been primarily confined to determining the presence or absence of risk aversion via a series of pairwise choices between a certain prospect and an uncertain one with equal expected value. These studies have shown clear, but far from universal, preference for the certain prospect under normal conditions (see, for example, Frank Logan, 1965; Leslie Real, 1981; J. E. R. Staddon and Nancy Innis, 1966; Keith Waddington et al., 1981; John Young, 1981).

We report three series of experiments here. In the first experiment, rats choose over prospects with equal expected values, which differ by mean preserving spreads (Michael Rothschild and Joseph Stiglitz, 1970). We test for risk aversion over the range of consumption variation covered, and transitivity of choices based on pairwise comparisons across prospects (weak stochastic transitivity). In addition, using a strength of preference measure, we determine if the preference ordering matches the ordering based on differences in risk premiums between prospects as expected utility theory suggests (strong stochastic transitivity). We find risk aversion (over food pellets) satisfied throughout, and weak, but not strong, stochastic transitivity over prospects. The second experiment looks at risk preferences at varying consumption levels, where consumption is measured in terms of total daily food intake. Consumption varies by a factor of 200 percent, with

*Texas A&M University, College Station, TX 77843; University of Houston, Houston, TX 77004; and Northeast Louisiana University, Monroe, LA 71201; respectively. Support from the National Science Foundation was essential to conducting this research. The advice of Len Green and the assistance of Carl Kogut and Steve White in conducting the experiments is gratefully acknowledged. Tom Caraco, Mark Machina, and Dan Levin's comments on earlier drafts helped clarify a number of conceptual issues. Responsibility for errors remains ours alone.

[1] Failure to maximize in discrete trial experiments only occurs when a correction or guidance procedure is used so that if subjects first choice on a trial is not reinforced on that trial, it then receives reinforcement on the other alternative (Sutherland-Mackintosh). The data from these experiments clearly support a maximizing, as opposed to a probability matching (R. Duncan Luce and Patrick Suppes, 1965), formulation.

low-consumption levels insufficient to cover long-run net energy requirements (and resulting in rapid and substantial weight losses), and with high-consumption levels approaching satiation. The rats display constant to mildly decreasing risk aversion throughout, with no tendency to risk loving under the negative net energy budget condition as some recent biological research suggests. The third experiment is designed to induce violations of the independence axiom of expected utility theory through mimicing conditions reliably resulting in violations with human subjects choosing over hypothetical outcomes. These involve tests of the "common ratio" effect (Maurice Allais, 1953), also called the "certainty effect" (Daniel Kahneman and Amos Tversky, 1979). Rats, too, violate the independence axiom in the Allais-type direction.

In a paper intended for an economics audience, no introduction would be complete without some brief remarks defending the use of animal subjects. This is particularly relevant in such a well-worked area of experimentation as expected utility theory. The primary advantage of animal subjects is methodological, as they involve a low-cost method of studying choices between real, highly valued alternatives, while most human experiments involve choices over hypothetical outcomes or choices involving relatively small payoffs. This gives our experiments heightened saliency (Vernon Smith, 1982), a common criticism of tests of expected utility theory which employ hypothetical or low stakes outcomes, particularly when the test results fail to support the theory (Mark Machina, 1983a). (Unfortunately, the data all too commonly show discrepancies between choices over hypothetical and actual payoffs.)[2] In addition, agents here receive a great deal of feedback regarding the outcomes of their choices, while studies involving human agents commonly employ little feedback. To the extent that one can argue that "rational" behavior is at least in part learned behavior generated as a consequence of interacting with one's environment, feedback is important.

Finally, as has been noted, outside of the extensive work on elements of first-degree stochastic dominance, there has been little research devoted to exploring questions of risk preference of animals of the sort typically of concern to economists.[3] As such, the experiments should prove of interest to those concerned with developing general descriptive models of choice in uncertain contexts, and of particular concern to those interested in developing reliable models of animal behavior in uncertain contexts; as in developing better, second-generation models, of optimal foraging behavior.

I. Experiment 1: Effects of Mean-Preserving Spreads on Choice

In this experiment, rats chose between alternatives that had the same expected value. Alternatives differed through mean-preserving increases in risk and could be ranked unambiguously as being more or less risky according to the Rothschild-Stiglitz definition. More precisely, subjects chose between

[2] For instance, Ebbe Ebbesen and Vladimir Konecne (1975) could not predict the bail set by judges in a courtroom from their choices when faced with hypothetical bail-setting problems in the laboratory. Richard Nisbett and Timothy Wilson (1977) found that verbal reports were often inconsistent with other measures of behavior, while Sidney Siegal (1961) found that the phenomena of probability matching, which is at odds with choice on the basis of first-degree stochastic dominance criteria, was very much dependent on the absence of financial payoffs. The literature is far from completely one-sided on the matter, however (Daryl Bem and Andrea Allen, 1974; David Grether and Charles Plott, 1979), indicating the need for testing propositions with real payoffs wherever possible.

[3] Psychologists have extensively compared choices between alternatives with a known time delay to reinforcement vs. an alternative with varying and uncertain time delays, where the mean time delays were the same (as was the size of the reward once delivered). This literature shows a marked preference for the alternative with uncertain time delays (Peter Killeen, 1968). This can be accounted for without invoking risk preferences in terms of a constant rate of time discounting, or a variable rate of time preference that overvalues more immediate rewards relative to more delayed outcomes, for which there is a considerable amount of independent evidence in the animal choice literature.

a pair of prospects, L and M, with the same means and with cumulative distributions, $L(x_i)$ and $M(x_i)$, with L larger than M in the sense that

$$(1)\quad \sum_{i=1}^{r} L(x_i)(x_{i+1}-x_i) \leq \sum_{i=1}^{r} M(x_i)(x_{i+1}-x_i)$$

$$r=1,2,\ldots,n-1; \qquad x_1<\cdots<x_n.$$

with strict inequality holding for at least one r. The paired prospects varied across experimental conditions, and could always be ranked as more or less risky according to (1).

If the rat's utility function is concave over the range of income variation involved (rats are risk averse), choice should always favor the less risky alternative. This in turn will assure transitivity of choice over the complete set of prospects. If the utility function is not globally concave, we will find choices favoring the more risky alternative under some paired comparisons. However, we can still test for transitivity of choices over the complete set of prospects. Intransitivities here would be inconsistent with von Neumann-Morgenstern utility theory as well as most (though not all) nonexpected utility models proposed in the literature (see Machina, 1983a, for a survey of model characteristics). These considerations give rise to our first two hypotheses.

HYPOTHESIS 1: *Utility functions are concave throughout the range of income variation so that in pairwise comparisons, choice always favors the less risky prospect.*

HYPOTHESIS 2: *Pairwise choices over the set of prospects* $\{A,B,C,\ldots\}$ *reveal no intransitive cycles.*

Measures for evaluating Hypotheses 1 and 2 consist of the relative frequency of choosing prospect L over M. In each case, the rat made a number of such choices. Choice was rarely, if ever, exclusively confined to one of the prospects, consistent with the overwhelming majority of data from comparable experiments involving both human (R. Duncan Luce and Patrick Suppes) and animal subjects (Caraco et al.; Real). (We deal more extensively with this aspect of choice in the discussion section.) As such we say that prospect L is preferred to M ($L \succ M$) if and only if $F(L,M)>.5$, where $F(L,M)$ is the relative frequency of choosing L over M in a pairwise comparison. In terms of the stochastic choice literature (Luce and Suppes), Hypothesis 2 involves a weak stochastic transitivity requirement.

In cases where choices over a set of prospects $\{A,B,C,\ldots\}$ shows risk aversion throughout (with lower letter alternatives associated with less-risky prospects), the risk premium increases monotonically across alternatives, with a larger difference in the risk premium associated with prospect C when paired with A, than when prospect B is paired with A, etc. If we assume that response frequency is sensitive to differences in risk premia, we should find prospect A chosen with the same or greater frequency when paired with a prospect with a larger risk premium. This suggests the following hypothesis:

HYPOTHESIS 3: *Between any set of prospects* $\{A,B,C\}$ *which satisfy Hypothesis* 1, *and can be rank ordered according to the Rothschild-Stiglitz criteria,*

$$(2)\quad F(A,C)\geq \min\{F(A,B),F(B,C)\}.$$

Condition (2) is referred to in the stochastic choice literature as a strong stochastic transitivity requirement. We look for strong stochastic transitivity under much more limited circumstances than is commonly done in the stochastic choice literature, out of consideration of the fact that between any arbitrary pair of prospects, L, M, utility differences are meaningless without much stronger assumptions on the structure of preferences than that imposed by expected utility theory (Luce and Suppes; Luce and Howard Raiffa, 1957, ch. 2). Further, we employ a weak, rather than a strong, inequality in (2) to account for cases where choice frequencies are always either zero or unity, as well as likely upper and lower

bounds on choice even when subjects do not display completely polar preferences. The latter may result from minimum sampling requirements necessary to track changes in environmental condition or other, longer-run, survival requirements.

A. *Experimental Procedures*

In all cases, rats chose between a single pair of prospects at a time (shown in Table 1). Choices were recorded and payoffs delivered in response to a single press on one of two choice levers. Each experimental session began with 8 forced-choice trials followed by a fixed number of free-choice trials (see Table 1). Trials were separated by a constant time interval of approximately 1 minute, irrespective of the choice made, the outcome obtained, or the prospects under consideration.

The forced-choice trials served to familiarize subjects with the alternatives. During these trials, only one of the choice levers was available. Further, the empirical distribution function was forced to match the programmed distribution function over each prospect's trial set. Sequences of choices across levers and outcomes were fixed; however, the start point was determined randomly on a daily basis.

The free-choice trials served to measure preferences. During these trials, subjects could choose which lever to respond on. For example, subject 210 in choosing over prospects A and D could devote all 22 of its free-choice trials to A, or all 22 to B, or pick any combination (and sequence) of A and D outcomes it preferred, subject to the constraint of 22 choices. The empirical distribution function was *not* constrained to match the programmed distribution function here: the random number generating algorithm was allowed free reign, and the probability of obtaining a given outcome on any trial was independent of outcomes on other trials.

Rats had no access to food between experimental sessions. They had a water tube available in the experimental chamber and were provided with ad lib access to water in their holding cages. Expected pellet payoffs were at a level that would assure the rat's health, but well below satiation levels. In all cases, the rats would have readily accepted more choice trials (see experiment 2 below).

TABLE 1—PROSPECTS USED IN EXPERIMENT 1

Prospect	P_{x_1}	x_1	P_{x_2}	x_2	Number of Free-Choice Trials
Subjects 210 and 211					
A	1.0	10			22
B	.5	4	.5	16	22
C	.5	1	.5	19	22
D	.75	1	.25	37	22
Subjects 303 and 323					
A'	1.0	8			17
B'	.5	4	.5	12	17
C'	.5	1	.5	15	17
D'	.75	1	.25	29	17

Note: P_{x_i} = probability of payoff x_i; x_i = number of pellets.

Experimental sessions were conducted once a day, 7 days a week, at approximately the same time each day. Rats chose over the same pair of prospects, with prospects fixed behind the same lever, for a minimum of 15 and a maximum of 24 days. Within this constraint, conditions were changed when visual inspection of the data indicated no trend in choices over a consecutive 5-day period. Conditions were commonly changed within 18 days.

Preference measures consist of the average proportion of choices of the less-risky alternative over the last 5 days of an experimental condition. To control for lever bias which can, at times, be severe for rats, we typically measured choice with prospects first on one lever (for example prospect A on the left lever, D on the right) and then switched prospects across the levers (D on the left, A on the right) and averaged the data for the last 5 days under each condition.[4] This amounts to assuming that the utility function is additively separable with respect to position bias and risk preferences.

[4]Position biases of the sort encountered here are not peculiar to rats but are found in humans (Howard Rachlin and Marvin Frankel, 1969) and other species, including insects (Real).

B. *Results*

Table 2 shows relative choice frequencies for the less-risky alternative, for the pairwise comparisons investigated. In each case we conducted two-tailed t-tests, based on daily data for the last 5 days of each condition, of the null hypothesis that the choice frequency was .50; that is, subjects were indifferent between the prospects under consideration. A replication is shown for subjects 303 and 323 of the A' vs. D' comparison that had taken place some six months earlier as part of their participation in experiment 2 reported below.

The data indicate that risk aversion is dominant in the rats as in virtually all cases the average choice of the more certain alternative exceeds 50 percent and commonly exceeds 60 percent. The only exception to this is subject 211 in choosing between prospects C and D, where prospect C, the less-risky alternative, is chosen less than 50 percent of the time. (Note that the t-test here indicates that we cannot reject the null hypothesis of indifference at the usual significance level.) Although choices far from exclusively favor the less-risky prospect, the t-statistics are generally different from zero using a 5 percent significance level criteria, indicating a clear preference on subjects part. Thus, the data are largely in agreement with Hypothesis 1, risk aversion over the complete range of income variation.

For subjects 210 and 211, in choosing between the A, B prospects, we simply indicate that A is preferred to B more than 50 percent of the time. In this one case we do not have a more precise measure since prospects were not switched across levers, but we simply compared B with A, with A on the right lever in both cases. Nevertheless, we infer preference for A over B, since this choice comparison followed a comparison of B with D, with D on the right lever, and resulted in both rats choosing prospect B over 70 percent of the time when paired with D, to less than 20 percent of the time when paired with A (compare Figure 1). With the positioning of prospect B held constant across conditions, position bias would have to be substantially more variable across prospects from anything we have seen for us not to conclude that $A \succ B$.

Acceptance of Hypothesis 1 implies that choices satisfy weak stochastic transitive over the several triads and the one quadrad we could compare directly. For both subjects 210 and 211, we have $A \succ B$, $B \succ D$, and $A \succ D$.[5] For both 303 and 323, we have two transitive triads each ($A' \succ B'$, $B' \succ D'$, and $A' \succ D'$; $B' \succ C'$, $C' \succ D'$, and $B' \succ D'$), and

TABLE 2—EFFECT OF MEAN-PRESERVING SPREADS: PERCENTAGE CHOICE OF MORE CERTAIN ALTERNATIVE[a]

	B	C	D
Subject 210			
A	>.5	X	61.4 (4.14)[d]
B	–	X	77.8 (12.20)[d]
C	–	–	65.0 (2.63)[c]
Subject 211			
A	>.5	X	82.7 (26.6)[d]
B	–	X	85.9 (11.45)[d]
C	–	–	45.9 (1.17)
Subject 303			
A'	52.9 (3.55)[d]	X	74.7[b] (6.64)[d]
B'	–	52.4 (.48)	78.2 (7.89)[d]
C'	–	–	53.5 (.86)
Subject 323			
A'	77.1 (4.11)[d]	X	86.5[b] (12.66)[d]
B'	–	64.7 (3.61)[d]	84.2 (7.91)[d]
C'	–	–	57.6 (1.46)

Note: t-statistics are shown in parentheses.
[a]Prospects listed in Table 1.
[b]Comparison of A' vs. D' some six months earlier for 303 and 323 show choice frequencies and t's of 64.7 (3.33)[d] and 68.3 (7.11)[d], respectively.
[c]Significant at 5 percent level.
[d]Significant at 1 percent level.

[5]The fact that $C \succ D$ or $C \sim D$ for 211 does not imply a violation of weak stochastic intransitivity as long as Hypothesis 1 holds for the remaining pairwise comparisons for this subject.

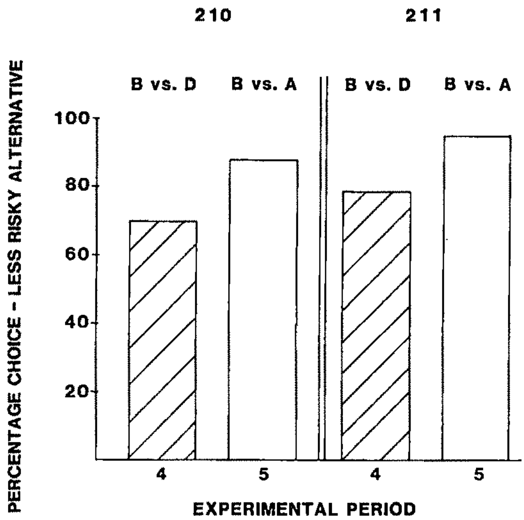


FIGURE 1. PERCENTAGE CHOICE LESS-RISKY ALTERNATIVE

(*Note*: Prospect on left choice lever listed first in each case.)

one transitive quadrad ($A' \succ B'$, $B' \succ C'$, $C' \succ D'$ and $A' \succ D'$).[6] The data clearly supports Hypothesis 2, weak stochastic transitivity. This is not to say that if we established prospects explicitly designed to generate intransitivities (see Tversky, 1969, for example), we might not find them. This is an empirical issue that we have yet to pursue.

Hypothesis 3 implies that moving from left to right along any given row of Table 2, we will find increasing frequency of choice for the less-risky alternative, as the differences in risk premia are increasing. Similarly, in moving down the columns of this table we should find decreasing frequency of choice for the less-risky prospect as the differences in risk premia are decreasing. Looking down the D column for subjects 210 and 211, we see that this prediction does not hold. Both subjects chose the less-risky alternative relatively more frequently when prospect B was paired with D, than when prospect A was paired with D. Ignoring the replication points for subjects 303 and 323 on the grounds that they were obtained at a substantially earlier time interval, we find similar results for 303 but not for 323. For subject 210, we reject the null hypothesis that $F(A,D) \geq F(B,D)$ at the 1 percent significance level ($t = 4.26$, $\alpha < .01$, two-tailed t-test). A combined evaluation of the t-statistics indicates that we can reject the null hypothesis $F(A,D) \geq F(B,D)$ at the 1 percent level also, primarily as a result of the outcome for subject 210.

Finding $F(B,D) \geq F(A,D)$, leads us to tentatively reject Hypothesis 3. Failure to support this hypothesis could be explained by several possibilities. In choosing prospect B (or B') with its moderate degree of uncertainty, relatively more than the sure alternative A (or A'), the rats could be exercising a preference for a moderate, controlled amount of uncertainty. Although this is incompatible with von Neumann-Morgenstern utility theory, a number of nonexpected utility models are compatible with this pattern of behavior (Clyde Coombs and Lily Huang, 1976; Machina, 1983b). Alternatively, one might argue that the rats, not knowing the prior distribution of the prospects under consideration, find they can afford to search more for large payoffs on the D (or D') prospect when they have the certain alternative to fall back on should they have bad luck. The B (B') prospect, with its moderate degree of uncertainty, does not provide a similar fall-back position, so the amount of search on the D (D') alternative must be reduced as a hedge against bad luck in the search process. The first explanation offered would be confirmed if we observed a preference for risk over prospects involving relatively small stakes or confirmed additional predictive implications of the nonexpected utility formulations which explain these data. We are currently conducting studies aimed at obtaining such observations.

C. *Discussion*

In choosing between prospects which differed by mean-preserving spreads, rats showed a consistent preference for the less-

[6]Paired comparisons of A' with C' for 303 and 323, which would have filled out the possible triad comparisons here, seemed like beating a dead horse at this point, so we devoted our scarce resources elsewhere.

risky outcome. This preference implies transitivity of choices in the sense of weak stochastic transitivity (Luce and Suppes). Assuming that choice frequencies vary monotonically with differences in risk premia, we found systematic inconsistencies relative to expected utility theory as three of four rats chose a moderately risky prospect relative to a high-risk alternative proportionately more than when choosing between a certain prospect and the high-risk alternative. One interpretation of these choice frequencies is that they represent a taste for a controlled amount of moderate risk. Note, however, that this preference did not extend to choosing moderately risky alternatives in favor of the sure outcome in paired comparisons. Experiments are currently underway to test the predictive consequences of this explanation.

Throughout the analysis we have dealt with choice frequencies averaged over the session as a whole, even though subjects made a number of choices during a given session. We have not bothered to distinguish the time frame for choice, whether it corresponds to a single trial, the session as a whole, or some number in between, in part because the predictions from expected utility theory hold irrespective of the time frame postulated. That is, any multiple of n choices of prospect D involves a mean-preserving spread relative to an equal number of choices of prospect A. However, as the animal progresses through a session, its deprivation level varies. This raises the obvious question of whether choice frequencies vary systematically over the course of a session.

To get a handle on this question we maintained sequential choice records throughout experiment 2 for six rats who were choosing between a certain outcome and an uncertain prospect over varying income levels. Two of these rats, 303 and 323, were involved in the present experiment, and one of the choice conditions in experiment 2 replicated one of the choice conditions involved here. Collecting sequential choice records from the last 5 days of each experimental condition, and dividing the free-choice trials into roughly equal halves, we tested for whether choices varied systematically between the first 8 and last 9 free-choice trials. We found no evidence for this in all conditions involving forced choice trials (for details see the results reported under experiment 2 below). Far less complete records of sequential choice patterns obtained from the present experiment support these results. Thus, variation in deprivation (or consumption) levels over the course of an experimental session exercised no systematic impact on choices.

Throughout our experiments, we find nonexclusive choices under conditions where expected utility theory suggests that risk-averse decision makers will exhibit exclusive preference for the more certain outcome. Nonexclusive choice here is consistent with the overwhelming majority of data from experiments involving human subjects choosing over hypothetical outcomes or small stakes prizes (see Luce and Suppes for a review), as well as the limited number of published studies on animal choices over uncertain outcomes (Real; Waddington et al.). It has provided the motivation for the development of stochastic choice theories (Luce and Suppes).

A number of possible explanations for nonexclusive choice suggest themselves in the context of the present experiments. One explanation that we can rule out is that it results from changing deprivation levels as the rats progress through the daily choice sequence. A number of alternative possibilities exist. Nonexclusive choice here is a direct prediction of nonexpected utility models incorporating a taste for a moderate, controlled amount of risk. In effect, the choice frequencies observed are a result of "mixing" prospects to obtain the most preferred level of risk (Coombs and Huang; Machina, 1983b). In addition, nonexclusive choice may result simply from search behavior on the rats' part. Recall that the rats' only knowledge of the relevant probabilities and payoffs comes from experiencing the outcomes. In experiencing uncertainty in nature, long-run survival would depend critically on monitoring the alternatives in the environment, and this undoubtedly transfers to the laboratory environment. Although the forced-choice trials were designed to refresh subjects' memories concerning the relevant payoffs and probabilities, this may not have proved sufficient to eliminate all search requirements.

Whatever the importance of this factor, we would expect the degree of search to be sensitive to its costs, so that we would find greater search the more nearly equal the prospects are in terms of their expected utility.[7] Further, for one reason or another, most rats display relatively strong side preferences, and we would not expect differences in risk between prospects to completely offset these biases. Here, too, however, one would expect actual choices to be sensitive to the costs of exercising these preferences. Finally, we probably have to accept some irreducible variability in choices, whether they be a result of minor variation in the environment or are elements of internal processes guiding an organism's choices.

II. Experiment 2: Effects of Varying Consumption Levels on Risk Preferences

In this experiment, rats again chose between prospects with equal expected values but with different degrees of risk (equation (1)). However, unlike experiment 1, the prospects under consideration remained constant while the rats' consumption level varied. The question of primary interest here is whether, and how, risk preferences vary with consumption level.

If choice frequencies are sensitive to the differences in risk premia associated with a given pair of prospects, then they provide direct measures of how the degree of absolute risk aversion (Kenneth Arrow, 1974; John Pratt, 1964) varies with consumption.[8] The conventional wisdom here argues for nonincreasing absolute risk aversion as consumption increases. This is the empirical pattern which is universally subscribed to with respect to data drawn from national economic systems (compare Machina, 1983a, for a survey of views on this point). This motivates our fourth hypothesis.

HYPOTHESIS 4: *Choice frequencies for the more certain prospect will remain constant or decline as consumption increases with the risk differential between prospects held constant.*

Recently a number of biologists have argued that risk loving is likely to be found at the very lowest consumption levels, when conditions reach the point where the expected value of the certain prospect falls below minimum survival requirements. The argument, roughly put, is that once consumption levels reach the point where an organism can no longer survive on the certain prospect, it actually increases its chances for survival playing the risky alternative and hoping to get lucky (see Caraco, 1981, and references cited therein). Caraco et al. report data from several experiments involving different species of small granivorous birds supporting this conjecture (also see Caraco, 1981). This motivates our next hypothesis.

HYPOTHESIS 5: *If consumption levels are sufficiently low that exclusive choice of the less-risky alternative would result in a negative net energy balance, and imperil the rat's survival, we will observe preference for the riskier alternative provided there is some positive probability that this choice will result in a positive net energy balance.*

Tests of Hypothesis 5 do not rely on choice frequencies measuring intensity of preference (differences in risk premia), but the weaker hypothesis that if one prospect is preferred to another, it will be chosen relatively more frequently. Results of tests of Hypothesis 5 may be of some relevance to understanding human behavior in low-income settings, or the likely effects of a firm's overall asset position (the "depth of its pockets") on its response to a continuing negative net income flow.

A. *Experimental Procedures*

Procedures were essentially the same as those employed in experiment 1 with the following modifications.

[7]See, for example, the interesting study by John Krebs et al. (1978) which tests for optimal sampling behavior on the part of birds.

[8]Caraco et al. found no difference in tests of changes in absolute risk aversion using a certainty equivalent measure of preference and the frequency of choice measure employed here (also see Caraco, 1982). The frequency of choice measure is substantially less time consuming to employ.

Subjects always chose between prospects consisting of a sure 8 pellets (prospect A' in Table 1) and a prospect delivering 1 pellet with probability .75 and 29 pellets with probability .25 (prospect C' in Table 1). There were 17 free-choice trials under all conditions. Two of the subjects, 303 and 323, participated in experiment 1, but the treatment conditions here actually preceded those reported in experiment 1.

Consumption levels were varied through adjusting the number of forced-choice trials across experimental conditions. These consisted of 32, 24, 8, and 0 forced-choice trials, and resulted in a 200 percent variation in consumption between the highest and lowest forced-choice conditions, had the rats allocated all their responses to the certain alternative. At the highest consumption level, conditions were approaching satiation for a limited duration choice session; in fact, two of the rats, 323 and 324, frequently failed to complete the free-choice trials within the allotted time interval. In these cases we measure preferences in terms of the percentage of free-choice trials completed, excluding days when less than 10 trials were completed. With zero forced-choice trials, the expected level of consumption was below minimum long-run survival requirements. All of the rats lost weight under this treatment condition, both at the beginning and the end, and were in clear danger of dying of starvation had conditions been maintained much longer.

B. *Results*

Table 3 shows the relative frequency with which the certain prospect was chosen as consumption (the number of forced-choice trials) varied. Also shown are t-tests of the null hypothesis of indifference (that choice frequencies are .50). With the exception of subject 323 under the 32 forced-choice trial condition, the certain prospect was chosen the majority of the time, and was chosen 60 percent or more in 17 out of 20 observations. Risk aversion holds up over virtually all consumption levels.

Average choice frequencies for the more certain prospect decrease somewhat in going from the 0 to 32 forced-choice condition (the mean at the 24 forced-choice condition is somewhat misleading due to the incomplete nature of our observations there). However, analysis of variance tests for systematic treatment effects across subjects indicates that this pattern is not statistically significant ($F = 1.24$, $d.f. = 3.16$, $\alpha = .33$). Eyeballing individual subject data, we find very little systematic variation in preference measures across individual subjects. In particular, with the notable exception of 323, no subject shows any clear, systematic decrease in preference for the risky alternative with increasing consumption levels as is required for decreasing absolute risk aversion. Further, of the two subjects who were not finishing regularly at the high consumption level, 323 shows risk neutrality but 334 shows relatively strong risk aversion.

Occasional subject idiosyncratic responses to changing consumption levels show up in Table 3, such as subject 303 under the 24 forced-choice trial condition and 334 under the 8 forced-trial condition. However, we do not attach much importance to such localized responses to the consumption variable, no matter what the degree of statistical significance, particularly if they are not replicated across subjects. For one thing, they are not suggestive of any regularities in behavior; for another, such effects are likely to be caused by transitory factors so that replication of the treatment condition would be unlikely to generate an equally strong response.[9]

The data in Table 3 clearly support Hypothesis 4, nonincreasing absolute risk aversion—not so Hypothesis 5, however. In no case do the rats prefer the risky prospect under the zero forced-trial condition, even though average weight dropped some 25 percent, from 250 to 180 grams, from the beginning to end of this treatment condition. Rats

[9] In a further effort to identify any systematic changes in individual subject risk preferences, we tried a number of covariance specifications, controlling for responses to good or bad luck consequent on choosing the risky prospect. Variables such as the difference between actual pellets received in a day less the expected number of pellets failed to provide any systematic explanation over choice frequencies, nor did various lagged specifications of this sort. As such, more formal analysis does nothing to change the conclusions reached in the text.

TABLE 3—FREQUENCY OF CHOOSING MORE CERTAIN ALTERNATIVE AT VARYING CONSUMPTION LEVELS

	Number of Forced-Choice Trials[a] (Total Pellets in Trials)				
Subject	32 (256)	24 (192)	8 (64)	0 (0)	Average Across Consumption Levels[b]
303	68.2 (4.87)[d]	82.4 (9.57)[d]	64.7 (3.93)[d]	63.0 (7.14)[d]	69.6 [7.62]
304	–	89.4 (11.32)[d]	90.0 (8.78)[d]	82.9 (20.48)[d]	87.4 [3.20]
323	45.9 (.44)	68.8 (4.44)[d]	68.3 (7.11)[d]	78.8 (3.51)[d]	65.4 [12.06]
324	74.1 (6.93)[d]	–	74.1 (6.11)[d]	80.6 (8.27)[d]	76.3 [3.04]
333	60.0 (5.11)[d]	–	67.1 (5.80)[d]	54.1 (1.72)	60.4 [5.31]
334	73.7 (3.42)[d]	–	54.7 (1.52)	76.5 (4.50)[d]	68.3 [9.65]
Average Across Subjects[b]	64.4 [10.5]	80.2 [8.5]	69.8 [10.7]	72.6 [10.5]	

[a] t-statistics are shown in parentheses.
[b] Standard errors are shown in brackets.
[c] Significant at the 5 percent level.
[d] Significant at the 1 percent level.

with ad lib access to food and water would weigh well over 400 grams at a comparable age, while at 180 grams we start to become seriously concerned with their survival. The data provide no hint of increasing risk preference here as for half the subjects, choices for the more certain alternative increase in going from 8 to 0 forced trials, while the average across rats increases. Finally, we found no shift in the data towards increasing risk preference during the last several days of this condition when the deleterious effects of these net energy deficits would have been at their maximum.

Sequential choice records, detailing the choice pattern within a daily session, were maintained throughout the experiment. We looked at these records over the last 5 days of each experimental condition to determine if there were any systematic temporal patterns to choice; for example, did choices of the risky alternative, D', tend to cluster towards the end of the free-choice trial period, or were they evenly distributed across an experimental session? We found two subjects, 304 and 324, who chose prospect D' relatively more frequently during the first half of the free-choice trials than would happen by chance alone (using either a 5 or 10 percent significance level criteria). For both these subjects, this uneven distribution of choices was confined to the zero forced-choice trial condition. Since these were the two most risk-averse (by the frequency of choice measure) rats under this treatment condition (as well as an average across all treatment conditions), only a handful of choices were allocated to the D' prospect. As such we interpret the uneven temporal distribution as search behavior induced by the absence of the forced-choice trials. Outside of this, there is no systematic tendency for choice distributions to vary as consumption and deprivation levels varied within experimental sessions.

Finally, since the data in Table 3 indicate approximately constant absolute risk aversion across consumption levels, while average choices of the certain alternative varied substantially across individual subjects, we could not resist determining whether there were significant individual subject differences in the level of risk aversion displayed. There are ($F = 3.14$, $d.f. = 5$, $\alpha = .04$). Differences in the structure of individual subject preferences have also been reported in consumer

choice and labor supply studies of animal behavior (Battalio et al., 1983), and in auction market experiments involving human subjects (James Cox et al., 1982). These results suggest a strong note of skepticism in evaluating economic models or empirical studies whose implications rely critically on the assumption of homogeneous preferences.[10]

C. *Discussion*

The present experiment tested for changing risk profiles under varying consumption levels. The rats all showed risk aversion, or risk neutrality, at all consumption levels with risk aversion clearly being the dominant response. This held even at the very lowest consumption level where net energy balance was not sufficient to insure long-run survival. The data support the hypothesis of nonincreasing absolute risk aversion for the group as a whole, with a slight tendency for reduced responding on the certain prospect between the lowest and highest consumption levels. Looking at individual subject data, only one rat (323) showed any systematic, statistically significant decrease in risk preference with increasing consumption.

We have found similar results in another experiment, virtually identical to the one reported here, in which rats chose over prospects consisting of water reinforcement, with ad lib access to food in their home cages (Kagel et al., 1984). Although with water reinforcement, the rats seem less risk averse in general (as judged by the frequency of choice measure). They show constant absolute risk aversion over a wide range of consumption variation and do not switch to risk-preferring choices at the lowest consumption level when faced with modest negative net energy balances. These results, of course, do not rule out decreasing risk preferences over even higher consumption levels, those that approach or equal ad lib consumption. In fact, pilot work in the water experiments suggests that this might well be the case, but we must reserve judgment on this point at the present time.

Our failure to find risk-loving behavior at the very lowest consumption level stands in marked contrast to the work of Caraco et al. There is no question that our rats could not meet long-run survival requirements under the zero forced-choice trial condition as indicated by both the change in, and absolute size of, their body weights.

There are a number of procedural differences between Caraco's experiments and ours that might account for the differences in experimental outcomes. Of greatest potential significance here is that the intertrial time interval covaried with changing consumption levels in Caraco's experiments, so that there was a substantially longer intertrial interval under the negative net energy budget conditions. In addition, the intertrial time interval varied directly with the outcome of the birds' choices, so that bad luck outcomes on the risky prospect were partially compensated for by an earlier start of the next choice trial, unlike the procedures employed here. This would interact with the changing average intertrial time interval to produce relatively greater partial compensation for bad luck under starvation conditions. Although we do not know the precise impact of these procedural differences on choices, we do know that there are circumstances where animals' choices are most sensitive to time delays associated with different prospects (see Kagel and Leonard Green, 1985, for a summary and discussion of the evidence on this point).

However, there are also significant differences in the weight of the birds Caraco used, approximately 20 grams normal weight, and their metabolism rates, as compared to the rats.[11] The upshot of this is that, other

[10] On this last point, we call special attention to Hans Binswanger's (1981) evaluation of the hypothesis of asset integration in uncertain contexts. Jaime Quizon et al. (1984) provide an alternative explanation for the failure of the asset integration hypothesis in Binswanger, as they note it is tested in conjunction with the premise that the utility function satisfies expected utility theory globally, and there is strong evidence that this auxilliary hypothesis is incorrect.

[11] Passerine birds such as the yellow-eyed junkos used in Caraco et al. have a metabolism rate which is some 65 percent higher than nonpasserine birds and mam-

things equal, the survival requirements of the birds are much more intense than the rats. For example, Caraco (personal communication) indicates that his birds could have lived approximately 3–4 days on the net energy flow associated with the expected value of the certain prospect employed, whereas our rats were able to live considerably longer. Further, it seems clear that with a relatively large mammal, such as rats, that we cannot induce such a relatively short time frame for survival when using food or fluid payoffs.

The basic idea underlying a switch from risk-averse to risk-preferring behavior as consumption levels change from an energy surplus to a deficit is that, with a deficit, the organism cannot meet post-foraging survival requirements through sticking with the certain alternative. As such, choosing the risky prospect actually increases the probability of survival. Large body weights relative to daily food requirements provide a buffer between short-term food intake and survival to the next foraging period as net energy deficits are met out of stored energy assets. The optimal decision rule is based on expected energy reserves, $E[S_t]$, relative to survival requirements, R. As long as $E[S_t] > R$, within a stationary environment, optimum behavior involves sticking with the certain alternative.

Relatively large mammals like rats facing negative net energy consumption levels are not likely to die of starvation per se, but rather of respiratory diseases induced by malnutrition. Thus, there might not be any clear-cut requirement R that needs to be met for survival, as in smaller organisms that are likely to simply drop from lack of energy and die from predation, so that the signal required for switching strategies is interferred with. In addition, predispositions to switching choices from risk averse to risk preferring, at least in foraging situations, might be much weaker in rats and larger organisms. The relative frequency with which a large organism (relative to daily food intake) would encounter such situations in nature in terms of evolutionary history, or the history of any individual organism, would have to be less than with relatively small organisms. Consequently, large animals might be much less sensitive to changing strategies in response to deteriorating environmental contingencies. Finally, the larger organism has much less to gain potentially by switching strategies, in terms of additions to stored energy assets, and may be better off as a rule to play it safe, avoid risks, and take its chances on the prospects for survival improving if and when foraging conditions change.

III. Experiment 3: Tests of the Independence Axiom

In verbal protocols with human subjects, cognitive psychologists and economists report systematic and replicable violations of the independence axiom of expected utility theory for large numbers of subjects. Of four types of systematic violations of the independence axiom reported in humans (Machina, 1983a), the "common ratio" effect (Allais) is most readily adapted to to our procedures (the common ratio effect includes the "certainty effect" of Kahneman and Tversky, and the "Bergen Paradox" of Ole Hagen, 1979, as special cases). The present experiment tests for this common violation of the independence axiom.

The common ratio effect involves rankings over pairs of prospects of the form:

Condition 1: Subjects choose between: A, with p chance of x_2, or $1-p$ chance of x_1 vs. B, with q chance of x_3, or $1-q$ chance of x_1.

Condition 2: Subjects choose between: C, with λp chance of x_2, or $1-\lambda p$ chance of x_1 vs. D, with λq chance of x_3, or $1-\lambda q$ chance of x_1,

where $p > q$, $x_3 > x_2 > x_1$, and $0 < \lambda < 1$ (the term "common ratio" derives from the equality of $\text{prob}(x_2)/\text{prob}(x_3)$ in A vs. B and C vs. D). An expected utility maximizer would prefer either A and C (if $p[U(x_2)-$

mals. C. J. Barnard and C. A. J. Brown (1985) have replicated Caraco's results using common shrews. Our remarks in the text apply equally to Barnard and Brown as they employed procedures similar to Caraco, and shrews are small bodied mammals (10 grams or so) living continuously on the edge of starvation.

$U(x_1)] > q[U(x_3) - U(x_1)])$, or else B and D (if the opposite were true). However, in verbal protocols researchers have found a systematic (although far from universal) tendency for subjects to prefer A and D. For example, Kahneman and Tversky (p. 266) found that 80 percent of their subjects preferred a sure 3,000 Israeli pounds (x_2) to an .80 chance of winning 4,000 Israeli pounds (x_3) or a .20 chance of winning nothing (x_1), while 65 percent preferred a .20 chance of winning 4,000 Israeli pounds or an .80 chance of nothing to a .25 percent chance of winning 3,000 Israeli pounds or a .75 percent chance of nothing.

Similar effects have been reported using a wide range of hypothetical payoffs from as high as \$1–\$5 million dollars in the original Allais formulation, to as little as \$100–\$200 in some recent reports by Kenneth MacCrimmon and Stig Larsson (1979). In some studies of our own, we found with hypothetical payoff values of $x_3 = \$20$, $x_2 = \$15$, $x_1 = \$0$, and $p = 1.0$, $q = .80$, and $\lambda = .25$, that 55 percent of our respondents (MBA students) preferred A under condition 1, while 65 percent preferred D under condition 2 (two-thirds of those choosing A, chose D in this sample).

Kahneman and Tversky explain these violations of the independence axiom in terms of a "certainty effect" which they incorporate into their theory via subjective probability assessments. Machina's (1982) explanation rests in large part on changing risk attitudes at different expected income levels associated with conditions 1 and 2.[12] These are but two of several alternative explanatory frameworks offered in the literature (see Machina, 1983a, for a recent review).

The present experiment tests for violations of the independence axiom under conditions where subjects received real, highly valued, payoffs and were thoroughly familiar with the outcomes of their choices; common criticisms of hypothetical choice protocols. This provides the basis for our last hypothesis.

HYPOTHESIS 6: *When faced with choices which mimic the "common ratio" effect typically found in verbal protocols, the majority of rats will violate the independence axiom in the Allais-type direction.*

A. *Experimental Procedures*

Experimental procedures were essentially the same as employed in experiments 1 and 2.

Table 4 shows the experimental treatment conditions. The probabilities of the 8-pellet payoffs in prospects *C* and *E* involve multiplying the probability in prospect *A* by .50 and .33, respectively. The probabilities for the 13-pellet payoff in prospects *D* and *F* are obtained in a similar fashion from the probability for that payoff in prospect *B*. The 1-pellet payoff serves to soak up the displaced probabilities in both cases. For reference purposes, we will refer to experimental conditions by the value of λ used in deriving one prospect from the other.

Also shown under each prospect in parentheses is the expected value, or actuarial value, of the outcomes. Note that the expected value of each outcome drops rather drastically as λ decreases. We decided to use a 1-pellet payoff rather than a zero-pellet payoff for the "unlucky" outcome out of consideration for these income effects.

B. *Results*

Table 5 shows the mean frequencies with which the lower expected value prospects (*A*, *C*, and *E*) were chosen at the various values of λ. The *t*-statistics testing the null hypothesis of indifference are shown in parentheses below these means with footnote keys c and d designating statistical significance.

At the $\lambda = 1$ value, all subjects display either a modest preference for the lower expected value prospect *A*, or indifference be-

[12] The force of Machina's argument would seem to be attenuated considerably for the low-income examples in MacCrimmon and Larsson, and the results of our verbal protocols, as differences in expected incomes between conditions would seem too small to plausibly account for most of the variation in preferences. However, the frequency with which violations of the independence axiom occur drops at these lower-income levels, as one would expect if Machina'a argument is correct.

TABLE 4—TREATMENT CONDITIONS EMPLOYED IN EXPERIMENT 3

Condition	Prospects			Value of λ
1	A: 8 pellets, prob. 1.0 (8.0)	or	B: 13 pellets, prob. 3/4 1 pellet, prob. 1/4 (10.0)	1.0
2	C: 8 pellets, prob. 1/2 1 pellet, prob. 1/2 (4.5)	or	D: 13 pellets, prob. 3/8 1 pellet, prob. 5/8 (5.5)	1/2
3	E: 8 pellets, prob. 1/3 1 pellet, prob. 2/3 (3.33)	or	F: 13 pellets, prob. 1/4 1 pellet, prob. 3/4 (4.0)	1/3

Note: Forced-choice trials = 16; free-choice trials = 20 under all conditions. The expected value is shown in parentheses.

TABLE 5—COMMON RATIO EFFECT:
FREQUENCY OF CHOOSING 8-PELLET PAYOFF ALTERNATIVE[a]

	Value of λ			
Subject	1.0	1/2	1/3	1.0[b]
304	57.0 (1.57)	55.0 (1.11)	46.0 (−.87)	56.0 (1.71)
324	47.8 (−.78)	41.0 (−2.94)[c]	36.0 (−4.00)[d]	–
333	61.0 (1.91)	54.0 (1.13)	–	–
334	65.5 (3.46)[d]	44.5 (−1.09)	47.5 (−.56)	69.0 (16.9)[d]
Average Across Subjects	57.8 [7.53]	48.6 [6.94]	43.2 [6.25]	62.5 [9.19]

Note: t-statistics are shown in parentheses; standard errors are in brackets.
[a]Subject 333 died before it could complete the $\lambda = 1/3$ treatment condition, while 324 died before we could replicate baseline conditions.
[b]Replication point following $\lambda = 1/3$ condition.
[c]Significant at 5 percent level.
[d]Significant at 1 percent level.

tween the prospects. Prospect B was, of course, chosen with this preference relationship in mind, since the standard violation of the independence axiom has $A \succ B$, or at least $A \sim B$. As λ decreases, there is a general erosion of preference for the 8-pellet payoff alternative. While three out of four rats preferred the sure thing alternative at $\lambda = 1.0$, three out of three prefer the prospect delivering 13 pellets when $\lambda = 1/3$ (the mean value here just approaches statistical significance in favor of prospect E, $t = 1.88$, 2 $d.f.$, $\alpha = .10$, one-tailed test). Further, behavior is reversible for the two rats, 304 and 334, for which we were able to replicate the $\lambda = 1.0$ condition, indicating a stable preference structure. The repeat reliability here is quite remarkable, even by the standards we have come to expect from the rats.

Although the swings in choice frequencies in Table 5 are far from overwhelming, they are consistent across rats and in the predicted direction. Further, two of the rats show a clear change in preference as λ varies. At $\lambda = 1.0$, 324 is indifferent between the prospects, but at $\lambda = 1/3$ clearly prefers prospect F to E, in violation of the independence axiom. At $\lambda = 1.0$, 334 has a clear preference

for prospect A, but, at $\lambda=1/2$ and $\lambda=1/3$, is indifferent with clear tendencies in favor of D and F, respectively. The data provide clear support for Hypothesis 6.

C. *Discussion*

Tests of the common ratio effect show clear violations of the independence axiom. The fact that subjects chose between real, highly valued, payoffs and were thoroughly familiar with the prospects under consideration, directly attacks one of the last lines of defense for the expected utility hypothesis, given the repeated violations found using hypothetical payoffs with human subjects. Extensions of the procedures employed here may be used to sort out between the various nonexpected utility formulations designed to explain these violations.

APPENDIX

A. *Equipment*

Experiments were conducted in a 32 cm. × 32 cm. × 30 cm. experimental chamber with a grid floor which was housed in a larger sound-insulated box. One wall of the chamber held two metal levers, one placed 6.5 cm. to the left and one to the right of the center of the wall, which extended 2.5 cm. into the chamber and were located 9.0 cm. above the floor of the chamber. Both levers required a force of approximately .15N for a response to result in reinforcement and to be recorded. Pellet dispensers designed to deliver 45 mg. Noyes food pellets delivered food into a trap at cage floor level, directly below each lever. Two 2–W white signal lights were mounted one on top of the other with the lower light 5 cm. above the left lever, with two 2–W red signal lights mounted in a similar fashion over the right lever. Two 2–W white signal lights were mounted on the ceiling and served as houselights. The wall opposite to the levers held a water bottle that was continuously available. White noise designed to mask extraneous sounds was continuously present and an exhaust fan provided ventilation. All experimental contingencies and data recording were controlled through the use of a Rockwell Aim 65 microprocessor.

B. *Procedures*

Each free-choice trial began with the illumination of both sets of lights over each lever. A single lever press resulted in reinforcement delivery and extinction of the lights over each lever. The houselights remained on during the reinforcement period. The reinforcement period was set equal to 2 sec. times the maximum number of pellets delivered in any trial over the complete set of prospects subjects faced. The length of the reinforcement period was the same irrespective of the alternative chosen or the outcome of the choice. A 30 sec. blackout period during which all lights were extinguished followed the reinforcement period, after which a new trial began with the houselights on and the lights lit over both levers. Forced-choice trials were identical to free-choice trials with the exception that the lights were only lit over the lever programmed for reinforcement. Sessions ended following completion of the predetermined number of free-choice trials or, in cases with substantial latency periods between choice, when the expected time interval of the session plus one-half hour had passed. Almost all sessions ended with subjects completing the full set of free-choice trials allotted.

Forced-choice trials were all in multiples of eight. For each set of eight, four were forced to the left-hand side and four to the right. Further, each set of four on the uncertain prospect were constrained to deliver the expected value. The start of the sequence varied randomly across sets of forced-choice trials. This forced subjects to exactly sample the relative frequencies of the different outcomes except in the case of prospect E in experiment 3, where we rounded the forced-choice payoffs to three 8-pellet outcomes (.375 vs. .33) and five 1-pellet outcomes (.625 vs. .667).

Sequencing of treatment conditions by subject across experiments are shown in Table A1.

Experimental sessions were conducted once a day, 7 days a week, at approximately

TABLE A1—TREATMENT SEQUENCES

Experiment 1[a]	
(prospect on left lever listed first)	
210; 211:	*D-A, A-D, D-A, B-D, B-A, D-B, D-C, C-D*
303:	*A′-B′, B′-A′, B′-D′, D′-B′, A′-B′, A′-D′, D′-A′, D′-C′, C′-D′, C′-B′, B′-C′*
323:	*A′-B′, B′-A′, D′-A′, A′-D′, C′-D′, D′-C′, B′-C′, C′-B′, B′-D′, D′-B′*
Experiment 2[b]	
(sequencing of forced-choice trials)	
303; 323:	24, 8, 0, 32
304:	24, 8, 0
324, 333, 334:	8, 32, 0
Experiment 3[c]	
(prospect on left lever listed first)	
304:	*B-A, A-B, D-C, C-D, E-F, F-E, B-A, A-B*
324:	*B-A, A-B, B-A, A-B, C-D, D-C, F-E, E-F*
333:	*B-A, A-B, D-C, C-D*
334:	*B-A, A-B, C-D, D-C, C-D, E-F, B-A, A-B, F-E*

[a]Subjects 303 and 323 completed the sequence in experiment 2 prior to experiment 1.
[b]Side switches occurred between changes in number of forced-choice trials.
[c]All subjects completed the sequence in experiment 2 prior to experiment 3.

the same time each day. Rats were housed in individual holding cages and only one rat was in an experimental chamber at any given time.

C. *Subjects*

Subjects 210 and 211 were male Spague-Dawley rats, approximately 6 months of age at the start of the experiment. All other rats were male Wistar strain, approximately 3 months of age at the start of the experiments. All subjects were experimentally naive.

REFERENCES

Allais, Maurice, "Le Comportement de l'Homme Rationnel devant le Risque, Critique des Postulats et Axiomes de l'Ecole Americaine," *Econometrica*, October 1953, *21*, 503–46.

Arrow, Kenneth J., *Essays in the Theory of Risk-Bearing*, Amsterdam: North-Holland, 1974.

Barnard, C. J. and Brown, C. A. J., "Risk-Sensitive Foraging in Common Shrews (*Sorex Araneus. L.*)," *Behavioral Ecology and Sociobiology*, forthcoming 1985.

Battalio, Raymond C., Dwyer, Gerald J. and Kagel, John H., "Tests of Some Alternative Theories of Individual Choice Behavior," University of Houston, September 1983.

Bem, Daryl J. and Allen, Andrea, "On Predicting Some of the People Some of the Time," *Psychological Review*, November 1974, *81*, 506–20.

Binswanger, Hans P., "Attitudes Toward Risk: Theoretical Implications of An Experiment in Rural India," *Economic Journal*, December 1981, *91*, 867–90.

Caraco, Thomas, "Energy Budgets, Risk and Foraging Preferences in Dark-Eyed Juncos (*junco hyemalis*)," *Behavioral Ecology and Sociobiology*, November 1981, *8*, 213–17.

______, "Aspects of Risk-Aversion in Foraging White-Crowned Sparrows," *Animal Behaviour*, August 1982, *30*, 719–27.

______, **Martindale, Steven and Whittman, Thomas S.,** "An Empirical Demonstration of Risk-Sensitive Foraging Preferences," *Animal Behaviour*, August 1980, *28*, 820–30.

Coombs, Clyde H. and Huang, Lily C., "Tests of the Betweeness Property of Expected Utility," *Journal of Mathematical Psychology*, June, 1976, *13*, 323–337.

Cox, James C., Smith, Vernon L. and Walker, James M., "Auction Market Theory of Heterogeneous Bidders," *Economic Letters*, 1982, *9*, 319–25.

Ebbesen, Ebbe B. and Konecni, Vladimir J., "Decision Making and Information Integration in the Courts: The Setting of Bail," *Journal of Personality and Social Psychology*, November 1975, *32*, 805–21.

Grether, David M. and Plott, Charles R., "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review*, September 1979, *62*, 623–38.

Hagen, Ole, "Towards a Positive Theory of Preferences Under Risk," in M. Allais and O. Hagen, eds., *Expected Utility Hypotheses and the Allais Paradox: Contemporary Discussions of Decisions under Uncertainty with Allais' Rejoinder*, Dordrecht: D. Reidel, 1979.

Kagel, John H. et al., "Effects of Varying Liquid Levels on Risk Preferences in Rats," University of Houston, 1984.

Kagel, John H. and Green, Leonard, "Intertemporal Choice Behavior: Evaluation of Economic and Psychological Models," in their *Advances in Behavioral Economics*, Vol. 1, Norwood: Ablex Publishing, forthcoming 1986.

Kahneman, Daniel and Tversky, Amos, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, March 1979, *47*, 263–91.

Killeen, Peter, "On the Measurement of Reinforcement Frequency in the Study of Preference," *Journal of Experimental Analysis of Behavior*, May 1968, *11*, 263–69.

Krebs, John R., Kacelnik, Alejandro and Taylor, Peter, "Tests of Optimal Sampling by Foraging Great Tits," *Nature*, September 1978, *275*, 27–31.

Logan, Frank A., "Decision Making by Rats: Uncertain Outcome Choices," *Journal of Comparative and Physiological Psychology*, 1965, *59*, 246–51.

Luce, R. Duncan and Suppes, Patrick, "Preference, Utility and Subjective Probability," in R. D. Luce et al., eds., *Handbook of Mathematical Psychology*, Vol. III, New York: Wiley & Sons, 1965.

______ **and Howard, Raiffa,** *Games and Decisions*, New York: Wiley & Sons, 1957.

MacCrimmon, Kenneth R. and Larsson, Stig, "Utility Theory: Axioms vs. Paradoxes," in M. Allais and O. Hagen, eds., *Expected Utility Hypothesis and the Allais Paradox*, Dordrecht: D. Reidel, 1979.

Machina, Mark J., "'Expected Utility' Analysis Without the Independence Axiom," *Econometrica*, March 1982, *50*, 277–324.

______, (1983a) "The Economic Theory of Individual Behavior Toward Risk: Theory Evidence and New Directions," Technical Report 433, Center for Research on Organizational Efficiency, Stanford University, 1983.

______, (1983b) "Stochastic Choice Functions Generated from Deterministic Preferences over Lotteries," mimeo., December 1983.

Nisbett, Richard E. and Wilson, Timothy DeCamp, "Telling More than We Can Know: Verbal Reports on Mental Processes," *Psychological Review*, May 1977, *84*, 231–59.

Pratt, John W., "Risk Aversion in the Small and in the Large," *Econometrica*, January-April 1964, 122–36.

Quizon, Jaime B., Binswanger, Hans P. and Machina, Mark J., "Attitudes Towards Risk: Further Remarks," *Economic Journal*, March 1984, *94*, 144–48.

Rachlin, Howard C. and Frankel, Marvin, "Choice, Rate of Response and Rate of Gambling," *Journal of Experimental Psychology*, June 1969, *80*, 444–49.

Real, Leslie A., "Uncertainty and Pollinator-Plant Interactions: The Foraging Behavior of Bees and Wasps on Artificial Flowers," *Ecology*, February 1981, *62*, 20–26.

Rothschild, Michael and Stiglitz, Joseph E., "Increasing Risk I: A Definition," *Journal of Economic Theory*, September 1970, *2*, 225–43.

Siegal, Sidney, "Decision Making and Learning Under Varying Conditions of Reinforcement," *Annuals of New York Academy of Science*, 1961, *89*, 766–83.

Smith, Vernon L., "Microeconomic Systems as an Experimental Science," *American Economic Review*, December 1982, *72*, 923–55.

Staddon, J. E. R. and Innis, Nancy K., "Preference for Fixed vs. Variable Amounts of Reward," *Psychonomic Science*, February 1966, *4*, 193–94.

Sutherland, Norman Stuart and Mackintosh, Nicholas John, *Mechanisms of Animal Discrimination Learning*, New York: Academic Press, 1971.

Tversky, Amos, "Intransitivity of Preferences," *Psychological Review*, January 1969, *76*, 31–48.

Waddington, Keith D., Allen, Tracy and Heinrich, Bernd, "Floral Preferences of Bumblebees (*Bombus Edwardii*) in Relationship to Intermittent versus Continuous Rewards," *Animal Behaviour*, August 1981, *29*, 779–84.

Young, John S., "Discrete-Trial Choice in Pigeons: Effects of Reinforcer Magnitude," *Journal of Experimental Analysis of Behavior*, January 1981, *35*, 23–29.

Unemployment Rate Dynamics and Persistent Unemployment under Rational Expectations

By MICHAEL R. DARBY, JOHN HALTIWANGER, AND MARK PLANT*

Macroeconomists typically build models of the determination of real output with unemployment rate movements explained, if at all, by an appended Okun's Law relationship. In this paper we show that this approach is misleading because it misses important sources of persistence in cyclical unemployment—and hence—real output. In particular, we show that unemployment rate dynamics imply "hump-shaped" cyclical unemployment characterized by persistence, even though unemployed workers have faulty information on wage rates for only a comparatively brief period of time. This result is based on the interaction of two stock or state variables—inventories and cyclically unemployed workers—which result in a brief economic shock having a prolonged effect on the economy.

In Section I we develop the partial-adjustment equation which governs the evolution of the unemployment rate. We show that if individuals differ significantly in their probability of leaving unemployment, unemployment dynamics imply a much slower recovery from recession than would be implied by probability values observed in normal times.

In Section II we develop new empirical measures of the fraction of the labor force who become unemployed each month, and the probability that an unemployed person will become employed or leave the labor force in a month's time. These measures and alternatives—equivalent if all individual probabilities are the same—are used to demonstrate that the unemployed differ substantially in their individual probabilities of leaving unemployment. We can broadly characterize the labor force as consisting of two groups. The first group consists of recent entrants into the labor market who are in the process of job shopping and other individuals who have selected careers in sectors (occupational or industrial) that involve low accumulation of specific capital. This group is characterized by high rates of entry into unemployment and high probability of leaving it; in normal times, the bulk of unemployment comes from this group. The second group consists of individuals with high degrees of specific human capital and "permanent" jobs; so they rarely become unemployed, but search for a long time to find a replacement job when they do become unemployed. Loss of permanent jobs will be more prevalent during recessions because firms in declining industries find it optimal to accelerate eventual reductions in their labor force at such times. The slow search process of this second group implies that they dominate cyclical unemployment during the recovery from recessions.

Further empirical work reported in Section III supports the following macroeconomic parable: money shocks do not directly effect the proximate determinates of unemployment but do so indirectly through inventories. High inventories (relative to sales) lead to abnormally and temporarily high rates of entry of both groups into unemployment. The high inventories also may mildly and temporarily depress the probability of leaving unemployment. The excess unemployment among the high-probability group is quickly eliminated, but those who have lost permanent jobs take many months to find a new job, even though each individual's

*Department of Economics, University of California, Los Angeles, CA 90024. Darby is also with the National Bureau of Economic Research. We acknowledge helpful comments from Sean Becketti, Robert Clower, Sebastian Edwards, Maxwell Fry, John J. McCall, members of the UCLA Money Workshop, seminar participants at the UCLA Institute of Industrial Relations, and anonymous referees. The latter Institute and the National Science Foundation have provided financial support for this research. Able research assistance was provided by Zaki Eusufzai, Joel Lander, and Maria Sison. Any opinions expressed are our own and not of any institutions with which we are affiliated; this is not a report of the National Bureau of Economic Research.

monthly probability of success is no lower than would be the case in normal times. Thus, individual maximizing behavior leads to substantially persistent effects on cyclical unemployment and output.

I. Unemployment Dynamics and Persistence

In this section we demonstrate that the logic of the standard search model implies persistent cyclical unemployment as a result of a single-period forecast error on the part of searching workers. In this theoretical exposition of unemployment dynamics, we abstract from the notion that some of the unemployed may not be searchers, but instead may be on temporary layoff, expecting recall. We do this in order to emphasize the separate role of search dynamics in yielding persistent effects from uncorrelated errors. Initially we assume that all workers are identical.

A. *Identical Workers*

The dynamics of the unemployment rate can be described by a simple discrete time model. Let u represent the unemployment rate, s be the number of unemployed searchers, and n be the labor force. By definition:

$$u = s/n. \tag{1}$$

It follows directly that

$$\Delta u = \frac{1}{n}\Delta s - \frac{1}{n}\frac{s_{-1}}{n_{-1}}\Delta n, \tag{2}$$

and, therefore,

$$\Delta u = \frac{1}{n}\Delta s - \frac{\gamma}{1+\gamma}u_{-1}, \tag{3}$$

where $\gamma = \Delta n / n_{-1}$ is the growth rate of the labor force. Define the search flow f as the rate (per period) at which people begin search, and let π be the probability per period that a searcher will become employed (or leave the labor force).[1] The change in searchers Δs is simply the inflow during the period less the outflow:

$$\Delta u = \frac{1}{n}(f - \pi s_{-1}) - \frac{\gamma}{1+\gamma}u_{-1}. \tag{4}$$

Note that $s_{-1}/n = u_{-1}n_{-1}/n = u_{-1}/(1+\gamma)$ and let $\phi = f/n$ denote the search rate (the fraction of the labor force beginning search in the current period); then

$$\Delta u = \phi - ((\pi+\gamma)/(1+\gamma))u_{-1}. \tag{5}$$

The growth-adjusted probability $\pi^* = (\pi+\gamma)/(1+\gamma)$ is dominated by π empirically, so we can interpret equation (5) as saying that the unemployment rate rises (or falls) as the search rate exceeds (is less than) the adjusted probability of employment times the lagged unemployment rate.

The natural unemployment rate $\bar{u}$ is found by setting $\Delta u = 0$ for the normal or long-run equilibrium values $\bar{\phi}$, $\bar{\pi}$, and $\bar{\gamma}$:

$$\bar{u} = \bar{\phi}((1+\bar{\gamma})/(\bar{\pi}+\bar{\gamma})) = \bar{\phi}/\bar{\pi}^*. \tag{6}$$

That is, the natural unemployment rate is the product of the normal search rate $\bar{\phi}$ and adjusted duration of search $(1/\bar{\pi}^*)$.[2] For example, suppose that under normal conditions, 2 percent of the labor force begins the search process each month, the probability of finding employment in a month is one-third (i.e., the expected duration of search is $1/\pi$ or 3 months), and the growth rate of the labor force is 0.002/month. Then the natural

[1] Kim Clark and Lawrence Summers (1979) and John Ries (1984) indicate that the probability of leaving the labor force (temporarily) is substantial for unemployed people in certain demographic groups—especially married women and teenagers.

[2] Actually, $\bar{\phi}$ and $\bar{\pi}$ are proportionally equal underestimates of the true continuous time rates because, in the former case, individuals who both enter and leave unemployment between surveys are omitted and, in the latter case, the effects of continuously compounded attrition. See fn. 17 below for details of the correspondence between continuous and discrete time measures.

unemployment rate is

$$\bar{u} = 0.020\frac{1.002}{0.3353} = 0.020 \times 2.988 = 0.0598.$$

This is close to the 6 percent rate which we would obtain by ignoring the growth adjustment and simply taking the ratio of $\bar{\phi}$ to $\bar{\pi}$.

Denote cyclical components with circumflexes so that $\hat{\phi} = \phi - \bar{\phi}$, $\hat{\pi}^* = \pi^* - \bar{\pi}^*$, and $\hat{u} = u - \bar{u}$. Then combining equations (5) and (6) and manipulating the result yields[3]

$$(7) \quad \Delta u = \hat{\phi} - \hat{\pi}^* u_{-1} + \bar{\pi}^*(\bar{u} - u_{-1}).$$

That is, an abnormally high rate of new searchers, or an abnormally low probability of finding an acceptable job, tends to increase the unemployment rate, but aside from these shocks the unemployment rate converges to its natural rate with a constant partial-adjustment factor $\bar{\pi}^*$.[4]

To illustrate the implications of equation (7), suppose for simplicity that an unexpected economic shock increases $\hat{\phi}$ and lowers $\hat{\pi}^*$ temporarily until information on the change in policy can be incorporated in the expectations of entrepreneurs, searchers, and the public. Figure 1 illustrates the behavior of the unemployment rate on the assumption that these expectations effects on search flow and the instantaneous probability of employment last for one period only. The figure shows that a one-period expectations error implies a persistent effect on unemployment, and hence on real output in the standard search-unemployment model. In the second period, the probability that each searcher finds a job is at the normal level, but the additional unemployment engendered in the first period takes time to work off. If we were to suppose that the initial effect of a restrictive demand shock is in part to build up inventories, and that these excessive inventories lead to $\phi > \bar{\phi}$ in successive periods, the interval in which unemployment rates rise would be prolonged.

It may seem strange at first that when workers realize that they made a mistake last period, π^* returns only to its normal value rather than going below $\bar{\pi}^*$ to compensate for the error. A bit of overshooting is possible if there were financing constraints, but, generally, past errors are forever bygones and optimal sequential search will imply a reservation wage which results in the normal probability of successful job search.[5] This is not to deny that the wage offer distribution is likely to shift over the course of the cycle. In particular, one would expect the mean of the offer distribution to fall as the economy moves into a recession. However, as demonstrated by Steven Lippman and John McCall (1976, 1985), a correctly perceived decline in the mean of the offer distribution will, in general, imply a decline in reservation wages but the effect on the probability

[3]A form of equation (7) which accounts for π and γ separately is

$$\Delta u = \hat{\phi} - \left(\frac{\pi}{1+\gamma} - \frac{\bar{\pi}}{1+\bar{\gamma}}\right)u_{-1} - \left(\frac{\gamma}{1+\gamma} - \frac{\bar{\gamma}}{1+\bar{\gamma}}\right)u_{-1}$$
$$+ \left(\frac{\bar{\pi}+\bar{\gamma}}{1+\bar{\gamma}}\right)(\bar{u} - u_{-1}).$$

This form will be used in the empirical work below. The negative sign of the $((\gamma/(1+\gamma)) - (\bar{\gamma}/(1+\bar{\gamma})))u_{-1}$ term appears puzzling because we naturally think of $\hat{\phi}$ increasing as $\hat{\gamma}$ increases. In a partial sense, however, the more of a given increase in search flow that comes from new entrants, the lower will be the unemployed *relative* to the labor force. In other words, a given increase in $\hat{\phi}$ causes a slightly smaller increase in Δu if it comes about from new entrants as compared to the previously employed because in the former case the denominator of the unemployment rate is increased as well as the numerator. This neglects any secondary effect on π should new entrants have an expected search duration which differs from that of the previously employed.

[4]The rate $\bar{\pi}^*$ is not necessarily constant over time. For example, in the empirical work we show that it varies with the demographic composition of the labor force. For our immediate purposes, however, we may assume that it is constant.

[5]See Steven Lippman and John McCall (1985). Simple sequential search models imply that the expected *remaining* duration of search is a constant $(1/\bar{\pi})$ for uncompleted spells of unemployment when workers correctly perceive the potential distribution of offers. Financing constraints could be introduced to make the reservation wage a decreasing function of the length of unemployment experienced.

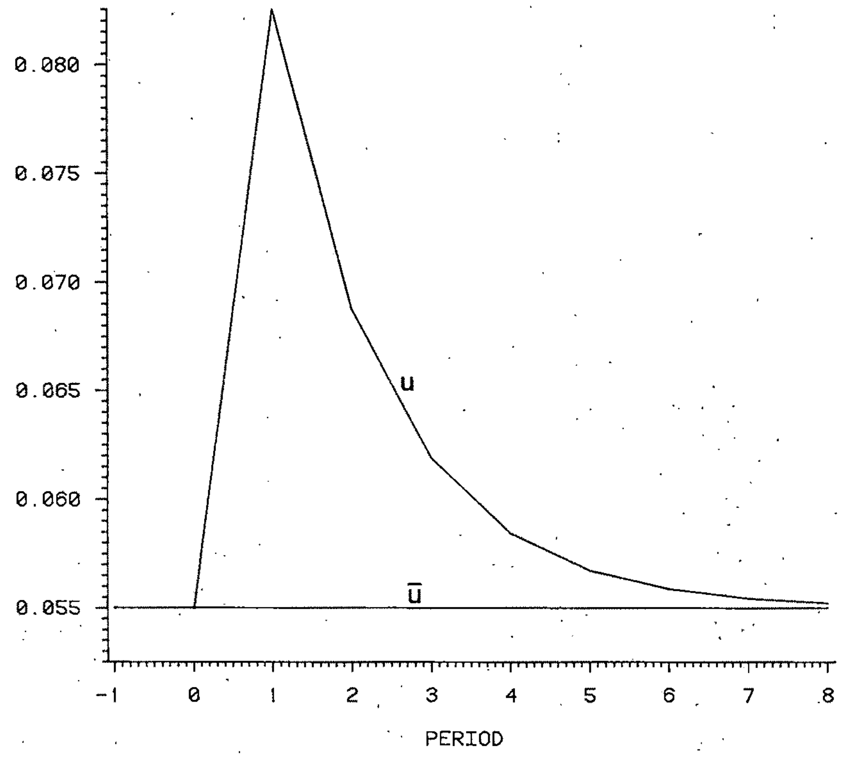


FIGURE 1. EFFECT ON U OF ONE-PERIOD SHOCK

of receiving an acceptable offer (π) will be ambiguous.[6]

The stock of search unemployed workers can be increased from its equilibrium value by a single-period expectational error. Once this excess stock comes into existence, a fraction is eliminated each period through successful search in the market place. Thus search dynamics alone imply persistent unemployment effects of one-time shocks in the economy. These effects are analogous to those previously discussed by equilibrium theorists for inventories, investment projects in process, and other such state variables. The persistent unemployment and output effects are not due to any persistent errors on the part of workers or firms, but instead due to the dynamic process of search in the labor market. In Section III below, we present empirical evidence that the interaction of inventory and unemployment dynamics does produce a hump-shaped time path of the unemployment rate in response to brief unexpected shocks.

B. *Heterogeneous Workers*

A second element of employment dynamics which plays an important role in explaining the persistence of unemployment

[6] It is worthwhile to remind the reader that the typical explanation of π being procyclical involves incorrectly perceived shifts in the offer distribution. For correctly perceived shifts, however, the effect on π is ambiguous as noted. It may be argued that the presence of positive or negative externalities associated with a change in the number of searchers (for example, Peter Diamond, 1982, and Peter Howitt, 1985) should play a role here. But these factors will involve a change in the shape (or a shift) in the offer distribution. Following the above arguments, if the change in the offer distribution is correctly perceived, then reservation wages are likely to change but any change in the probability of accepting an offer is likely to be second-order in magnitude.

subsequent to a macroeconomic shock is the heterogeneous nature of workers. As a general proposition, it has become increasingly accepted that heterogeneity across workers is important for understanding the behavior of unemployment. Realization that workers in different submarkets (for example, demographic, industrial, occupational) may exhibit different turnover behavior has led to a recognition that the natural rate of unemployment will vary with the composition of the labor market (see, for example, Michael Wachter, 1976; David Lilien, 1982; Haltiwanger and Plant, 1984).[7] Heterogeneity has also played an important role in the study of the distribution of unemployment duration. Specifically, the presence of heterogeneity implies a bias in using the observed (from the BLS household survey) uncompleted spell distribution to measure the unobserved completed spell distribution (see, for example, Stephen Salant, 1977; George Akerlof and Brian Main, 1980). Thus, we are clearly not the first to identify heterogeneity as an important factor for understanding the labor market. Rather, our innovation is to examine the ramifications of heterogeneity for persistence in unemployment over the business cycle. In particular, in this section we demonstrate theoretically how heterogeneity is potentially a source of persistence over the cycle, and then in succeeding sections we provide empirical evidence that indicates that heterogeneity is a major factor underlying the observed persistence in π (and hence in u).

For simplicity, we can think of workers as being divided into two groups. Those in the first group have little firm-specific human capital and they experience unemployment frequently, but the length of these spells are brief. Thus, ϕ_1 and π_1^* are large since there is little to be gained from extensive search for short-term employment. Members of the second group rarely experience unemployment, but when it occurs, search is extensive and well supported by unemployment compensation, other family income, and assets. So ϕ_2 and π_2^* are both low.[8] The normal unemployment rate is

$$(8) \quad \bar{u} = \frac{n_1}{n}\bar{u}_1 + \frac{n_2}{n}\bar{u}_2 = \frac{n_1}{n}\frac{\bar{\phi}_1}{\bar{\pi}_1^*} + \frac{n_2}{n}\frac{\bar{\phi}_2}{\bar{\pi}_2^*}.$$

We cannot observe individual values of the π_i's and ϕ_i's, but only their appropriately weighted averages. Of particular interest is π:

$$(9) \quad \pi = \sum_i \frac{s_{i,-1}}{s_{-1}}\pi_i = \pi_2 + (s_{1,-1}/s_{-1})(\pi_1 - \pi_2).$$

This overall average probability of leaving unemployment can change either because of changes in the individual π_i's or because of changes in the unemployment shares (s_i/s):

$$(10) \quad \Delta\pi = \sum_i \frac{s_{i,-2}}{s_{-2}}\Delta\pi_i + \sum_i \pi_i \Delta\left(\frac{s_{i,-1}}{s_{-1}}\right).$$

Consider once again a one-period shock. To concentrate on issues of persistence, suppose that people always search optimally ($\pi_i = \bar{\pi}_i$ always), but that a one-period increase in the ϕ_i's results in the same proportionate increase in each s_i. In the two-group case, $s_2/s = 1-(s_1/s)$. Holding π_1 and π_2 constant, we have from equation (10):

$$(11) \quad \Delta\pi = (\bar{\pi}_1 - \bar{\pi}_2)\Delta(s_{1,-1}/s_{-1}).$$

For the first two periods, by assumption, lagged s_1/s is unchanged so the observed π

[7]We incorporate this idea into our measure of the natural rate of unemployment in Section II.

[8]We have analyzed multiple groups of workers, but two groups are sufficient to capture the most important empirical features. We here explain the tendency, *ceteris paribus*, for π to decline as duration of unemployment increases by the sorting hypothesis: The expected duration of unemployment varies inversely with π, so that low π individuals comprise a larger share of longer duration relative to short duration unemployment. See James Heckman and George Borjas (1980), and John Carlson and Michael Horrigan (1983). Alternatively, π might itself decrease as duration increased. As discussed below, we believe that the data are best explained by the sorting hypothesis.

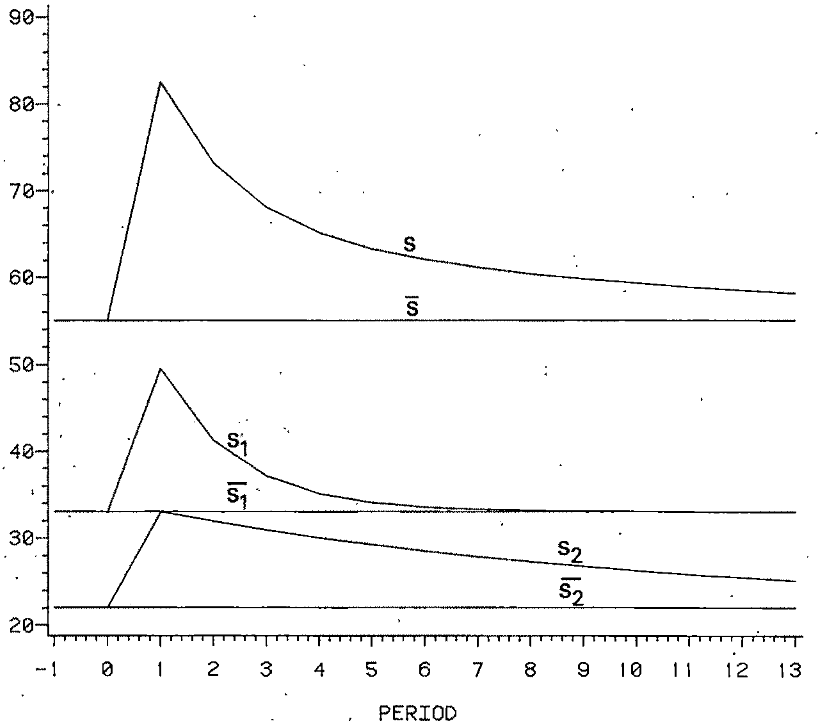


FIGURE 2. EFFECT ON s, s_1, s_2 OF ONE-PERIOD SHOCK

remains constant at $\bar{\pi}$. For each group, unemployment rates will follow a pattern like that exhibited in Figure 1, but reflecting the $\bar{\pi}_i$ appropriate to the group. In Figure 2 we plot the number of persons unemployed in each group, s_i, which is simply the product of the group unemployment rate and the size of the group ($s_i = u_i n_i$).[9] Once the search flows return to normal, the excess unemployment is eliminated in each group at the rate $\bar{\pi}_i$. At time 2, the lagged values of s_1 and s_2 have increased by the same proportion so (by equation (11)) π is unchanged at $\bar{\pi}$. But π thereafter begins to decline for a number of periods as the share s_1/s drops. This happens because $\pi_1 > \pi_2$, so s_1 returns to normal much faster than does s_2. Eventually, however (after about period 6 in the figure), nearly all the adjustment in group 1 is completed so that the decreases in s_2 become proportionately larger. Thereafter, s_1/s and hence π rise back toward their normal levels.

In conclusion, the slower adjustment of the lower-probability group will appear in the aggregate data as a persistent $\hat{\pi}$ even though each individual worker is searching optimally with $\pi_i = \bar{\pi}_i$. The problem arises because $\bar{\pi}$ is based on the normal distribution of unemployment among groups:

$$(12) \qquad \bar{\pi} = \frac{n_1 \bar{u}_1}{n\bar{u}} \bar{\pi}_1 + \frac{n_2 \bar{u}_2}{n\bar{u}} \bar{\pi}_2.$$

After a recession causes mass disemployment, the low-probability group will be overrepresented for a considerable period of time. During the recovery period (with $\phi_1 = \bar{\phi}_1$, $\phi_2 = \bar{\phi}_2$, and $\gamma_1 = \gamma_2 = \bar{\gamma}$), u_1 quickly returns to $\bar{u}_1$ and thereafter we observe

$$(13) \qquad \Delta u = \bar{\pi}_2^* (\bar{u} - u_{-1}).$$

[9] Since we have neglected the growth in the labor force, the transformation from u to s only shifts the relative positions of u_1 and u_2 so that they can be added together to obtain $u = (n_1/n)u_1 + (n_2/n)u_2$.

Were $\bar{\pi}_2$ sufficiently low, the return to the normal unemployment rate could be painfully slow in the absence of a later stimulative monetary or fiscal shock.[10] Note also that during this long recovery period the share of group 2 unemployed will be abnormally large which implies from equation (9) that π will be abnormally low ($\hat{\pi} < 0$). As a result, an analyst looking at only the aggregate equation (7) might incorrectly conclude that convergence to the natural unemployment rate $\bar{u}$ would be much faster were persistent expectational errors not keeping $\hat{\pi}$ negative. But we are considering a case in which each individual always correctly perceives the wage distribution, so that $\pi_i = \bar{\pi}_i$ always. So persistence in aggregate $\hat{\pi}$ may reflect expectational error or significant heterogeneity in π across individuals.[11]

This discussion of heterogeneity suggests that considerable care must be taken in defining the natural rate of unemployment. Movements in the unemployment rate will be highly correlated with changes in the composition of the pool of unemployed workers. So any measure of the natural rate that uses weighted averages of normal levels of π_i and ϕ_i where weights are unemployment shares (for example, $\bar{\phi}/\Sigma_i(s_i/s)\bar{\pi}_i$) will result in most movements in u being explained by movements in the natural rate.[12] Adjustment to normal levels of unemployment is reflected in adjustments in the share of unemployment among groups, and thus variations in shares should not be used to capture variation in the natural rate of unemployment.

II. Measurement

We demonstrated that the determinants of search rate ϕ, the employment probability π, and the labor force growth rate γ were important in the analysis of the cyclical pattern of unemployment and real output. In this section we first consider the measurement of ϕ, π, and γ, then decompose the measured values into cyclical and normal components, and finally obtain a measure of the heterogeneity of π across individuals. At present, we are not able to measure (separately) values of ϕ and π for those on temporary layoffs and those who are searchers (all others).[13] We postpone the development of such measures to future research.

A. *The Measurement of ϕ, π, and γ*

Since the size of the labor force is a regularly reported statistic, we have no difficulty in computing γ, the growth rate of the labor force. To measure ϕ and π, we would ideally like to have data that reports the gross flows of the number of workers among three states: employed, unemployed, and not in the labor force. Unfortunately, complete data of this sort are not regularly and reliably collected.[14] However, we can obtain very good estimates of π and ϕ from the available data on the aggregate number, s, unemployed each month and the number, s^{0-4}, who have been unemployed "0–4 weeks."

Recall that the *Current Population Survey* is conducted each month during the week which contains the twelfth day of the month. Thus the typical year has twelve surveys of which eight are conducted 4 weeks subsequent to the previous survey, and four are 5

[10] This model would appear to provide a new basis for Axel Leijonhufvud's corridor notion (1981) in which the economy converges to long-run equilibrium abnormally slowly after a major perturbation. Note, however, that it is hard to believe that exceedingly low values for $\bar{\pi}_2$ could be socially optimal even if they were privately optimal given our system of transfer payments. It is not clear—of course—that stimulative monetary or fiscal policy has the same power to reemploy group 2 workers as it does to disemploy them through bankruptcies and permanent layoffs.

[11] Recall that in the case just considered the accounting identity (7) will continue to hold although equation (13) is governing the convergence to the natural rate.

[12] Some hard-won lessons in this regard are imparted in Section III below.

[13] If we could, we would like to measure ϕ_L as the fraction of the labor force beginning temporary layoff within the month, and π_L as the probability of being recalled from temporary layoff within the month.

[14] For a discussion of the availability of and problems with the gross flow labor force data, see R. Smith and J. Vanski (1979).

weeks subsequent to the previous survey. We find it convenient to use standard months of 4.35 weeks (30.4 days = 365/12). Since s^{0-4} reports the number of people unemployed 31 days or less, it is a good measure of the people who have become unemployed over the last month.[15] The number unemployed at the last survey is used to approximate s_{-1}, the number unemployed exactly one month earlier. The equivalence is exact whenever u_{-1} equals the value $\phi(1+\gamma)/(\pi+\gamma)$ toward which the unemployment rate currently converges as it would, for example, in steady-state equilibrium; otherwise the error introduced in estimating π is bounded by 0.02 for plausible values of the relevant parameters.[16] These data are sufficient to calculate π as

$$\pi = 1-(s-s^{0-4})/s_{-1}, \tag{14}$$

since $1-\pi$ is the fraction of individuals unemployed last month who are still unemployed.

Our measured π tells us how much longer the average currently unemployed person would be unemployed under current conditions: $(1/\pi)$ months. This would be the expected total duration of the average newly unemployed person only if the ratio of search flows, f_1/f_2, equals the ratio of the unemployment shares, s_1/s_2. In steady-state equilibrium, for example, f_1/f_2 exceeds s_1/s_2 so that the average expected duration of newly begun unemployment is less than the average expected *remaining* duration of the currently unemployed.

The corresponding measure of ϕ is

$$\phi = s^{0-4}/n, \tag{15}$$

where n is the civilian labor force. Note that this discrete time measure of ϕ measures the flow of people who become unemployed between monthly observations and are still unemployed at this monthly observation. The continuous-time search rate would be higher.[17]

Monthly values of our estimates of ϕ and π are available from the authors on request. Figure 3 plots quarterly averages of the monthly values of π. The mean value of 0.46 is interpreted as saying that on average 46 percent of the people who are unemployed at the beginning of a month will find a job, leave the labor force, or be recalled from layoff by the end of the month. The vertical lines mark cyclical peaks (P) and troughs (T) on the NBER reference cycle chronology. We note first that the values of π tended to be rather higher (around 0.6) during the Korean and Vietnamese war eras, and also that the monthly probability of getting a job drops sharply during a recession and rises in a boom. Our estimate of $\bar{\pi}$ is derived and discussed in Section II, Part B, below.

[15] Census enumerators, who have a calendar before them during an interview, round unemployment duration to the nearest whole number of weeks. Correction for the difference between 30.4 and 31 days is not attempted.

[16] Since π normally lies between 0.4 and 0.6, this is an acceptable margin of error. See the Appendix, Part A, for details.

[17] Let ψ be the continuous-time search rate, θ be the Poisson parameter such that the probability of finding a job between t and $t+dt$ is $\theta\,dt$, and $\bar{\gamma}$ be continuously compounded monthly growth rate of n. Then the observed value of ϕ is given by

$$\phi = \left(\int_0^1 n_1 e^{\bar{\gamma}(1-t)}\psi e^{-\theta t}\,dt\right)/n$$
$$= \psi(1-e^{-\theta-\bar{\gamma}})/(\theta+\bar{\gamma}).$$

Note that the observed value of π is simply $1-e^{-\theta}$. Taking $\bar{\gamma}=0$, the following correspondence are observed:

π	θ	$\phi/\psi=\pi/\theta$
0.2	.223	.896
0.3	.357	.841
0.4	.511	.783
0.5	.693	.721
0.6	.916	.655

Thus the observed monthly probability of finding a job π is less than its continuous-time equivalent θ, and the observed monthly flow into unemployment ϕ is less than the continuous-time equivalent ψ.

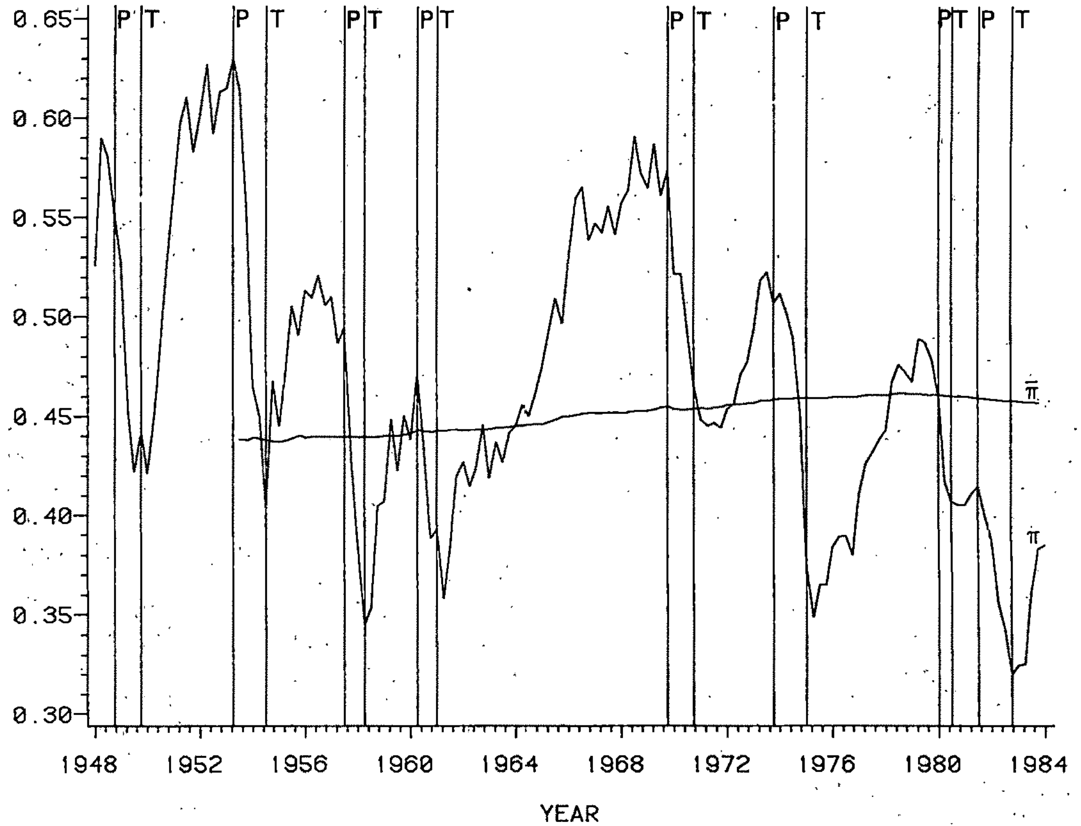


FIGURE 3. PROBABILITY OF EMPLOYMENT

Figure 4 similarly plots quarterly averages of monthly values of ϕ. We see that ϕ generally follows an upward trend although it is sharply below trend in both the Korean and Vietnamese war eras. The cyclical behavior is sharply contracyclical—rising in recessions and falling in booms.

B. *Cyclical and Normal Components*

In equations (6) and (7) we analyze unemployment rate dynamics in terms of normal values $\bar{\phi}$, $\bar{\pi}$, $\bar{\gamma}$, and $\bar{u}$ and the corresponding cyclical components $\hat{\phi}$, $\hat{\pi}$, $\hat{\gamma}$, and $\hat{u}$. In this section we develop estimates of these quantities. Following Robert Barro (1977, 1978) and Wachter, we develop measures of $\bar{\pi}$, $\bar{\phi}$, and $\bar{\gamma}$ which reflect the effects of the military draft and the age-sex composition of the labor force on the natural rate of unemployment and derive a measure of $\bar{u}$.[18]

[18]It might be useful to examine the effect of the industrial or occupational composition of the labor market on $\bar{\pi}$, $\bar{\phi}$, and $\bar{\gamma}$. This is of interest because the recent work of Lilien suggests that the industrial composition is important for measuring the overall natural rate of unemployment. However, to incorporate these ideas in this context, unemployment duration data broken down simultaneously by age, sex, industry, and occupation would be required, and such a four-way classification of this data is not available. Confronted with the choice of which decomposition matters most, we chose the age-sex classification. This is because the age-sex breakdown of the unemployment duration data is relatively more available and previous studies (for example, Wachter; Haltiwanger-Plant) have demonstrated that the age-sex composition of the labor force is very important.

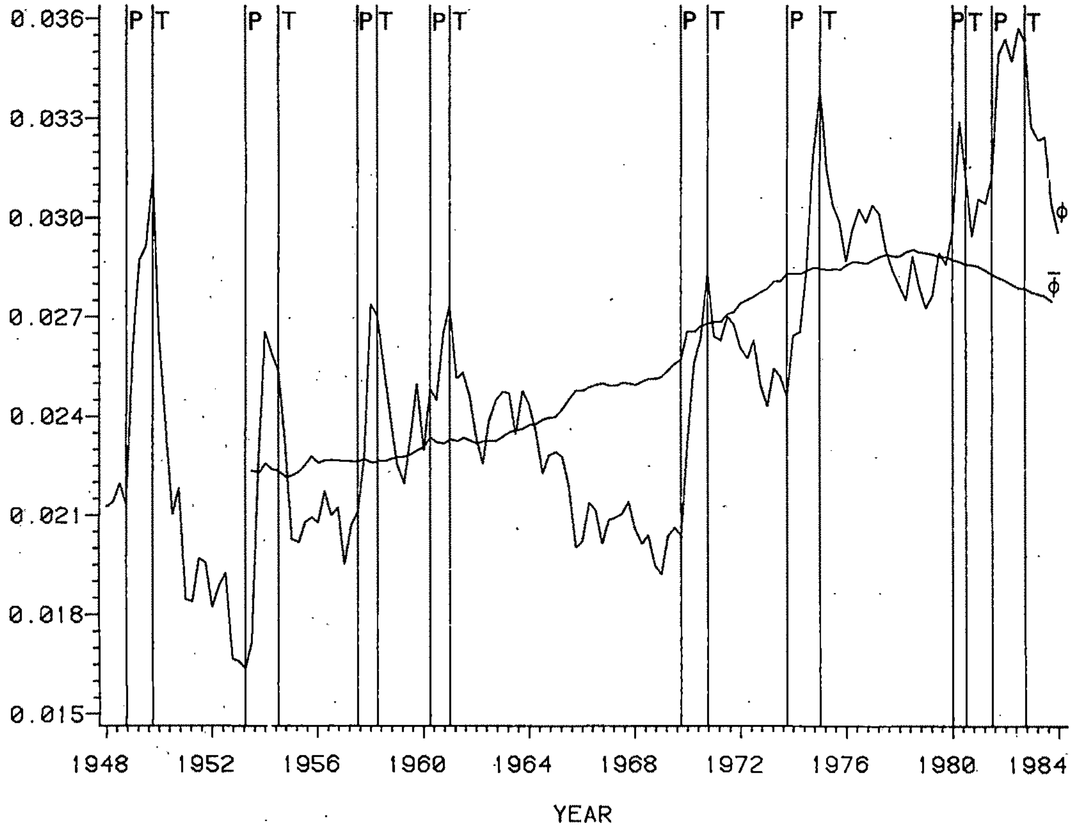


FIGURE 4. NET SEARCH RATE

We define $\bar{\pi}$ by

$$(16)\qquad \bar{\pi}=\sum_i \frac{n_{i,-1}\bar{u}_{i,-1}}{\sum_j n_{j,-1}\bar{u}_{j,-1}}\bar{\pi}_i,$$

where the summations are over age-sex cells.[19] Labor force shares and unemployment rates by age and sex are available monthly, but the number who have become unemployed during the last month by age and sex is only available as an annual monthly average beginning in 1967.

For all age and sex groups other than males 16–19 and 20–24, we hypothesize that $\bar{\pi}_i$ is constant over time. This allows us to use the available data to measure $\bar{\pi}_i$ with the following steady-state version of equation (14) for group i where $A(\ )$ denotes annual averages:[20]

$$(17)\qquad \bar{\pi}_i=1-\text{mean}\left[\frac{A(s_i)-A(s_i^{0-4})}{A(s_{i,-1})}\right].$$

[19] We must be careful to use normal unemployment weights in formula (16) since we have seen that after a recession causes mass disemployment, low-probability groups—predominant among prime-age males—will be overrepresented for some time. Therefore the use of actual unemployment weights will lead to spurious procyclical movements in measured $\bar{\pi}$ and hence contracyclical movements in $\bar{u}$. This type of problem seems to be at the root of other estimates of $\bar{u}$ which fluctuate sharply with the business cycle.

[20] All measures of normal $\bar{\pi}$, $\bar{\phi}$, $\bar{\gamma}$, and $\bar{u}$ are based on data from 1967 to 1983. The $A(s_{i,-1})$ denotes the annual average of data from the prior December through November of the current year, and the mean is taken over the period 1967–83.

For males 16–19 and 20–24, the normal $\bar{\pi}_i$ is allowed to vary over time with the nature of the military draft. To measure $\bar{\pi}_i$ for these two young male groups, we estimate the following equation:[21]

$$(18)\quad A(\pi_i)=\alpha_0+\alpha_1 A(MIL) + \alpha_2 A(\pi_j-\bar{\pi}_j)+\alpha_3 A(\pi_k-\bar{\pi}_k),$$

where π_i refers to either males 16–19 or 20–24, MIL is Barro's military draft variable, and $\pi_j-\bar{\pi}_j$ and $\pi_k-\bar{\pi}_k$ are the deviations from the mean of the annual average of monthly π_j and π_k for males 35–44 and 45–54, respectively.[22] The latter two variables are included to account for cyclical variation in π_i. Using the estimated coefficients from this equation, $\bar{\pi}_i$ for males 16–19 and 20–24 is calculated as

$$(19)\quad \bar{\pi}_i=\hat{\alpha}_0+\hat{\alpha}_1 MIL.$$

We define $\bar{\phi}$ as

$$(20)\quad \bar{\phi}=\sum_i \frac{n_i}{n}\bar{\phi}_i,$$

where the summations are over age-sex cells.

Similar to our measurement of $\bar{\pi}_i$, the normal $\bar{\phi}_i$ for group i for age-sex groups other than males 16–19 and 20–24 is measured by the 1967–83 mean of equation (14) for group i:

$$(21)\quad \bar{\phi}_i=\text{mean}\left[A\left(s_i^{0-4}\right)/A(n_i)\right].$$

For males 16–19 and 20–24, $\bar{\phi}_i$ is measured using an adjustment for the effect of the military draft similar to that used to adjust $\bar{\pi}_i$ for the military draft.

We define $\bar{\gamma}$ as

$$(22)\quad \bar{\gamma}=\sum_i \frac{n_i}{n}\bar{\gamma}_i,$$

where the summations are over age-sex cells and $\bar{\gamma}_i$ is measured as the mean of the observed monthly growth rates for group i.

Having measured $\bar{\pi}_i$, $\bar{\phi}_i$, and $\bar{\gamma}_i$ for each group i, we can measure $\bar{u}_i$ by an age-sex specific version of equation (6). The aggregate values $\bar{\pi}$ and $\bar{\phi}$ are obtained from equations (16) and (20), and used with $\bar{\gamma}$ and equation (6) to obtain the aggregate normal unemployment rate $\bar{u}$ which rises (see Figure 5) until the late 1970's and then begins to drift downward.

Examining Figures 3 and 4, we see that $\bar{\pi}$ is relatively constant and the key factor moving $\bar{u}$ is $\bar{\phi}$, which indeed increases during the 1970's and then falls off somewhat in the early 1980's. This can be explained by examining the estimates of the $\bar{\pi}_i$ and $\bar{\phi}_i$ values.

Table 1 reports the estimated normal values for $\bar{\pi}_i$, $\bar{\phi}_i$, $\bar{\gamma}_i$, and $\bar{u}_i$ by age and sex.[23] During the 1970's, the labor force share of the young increased and then decreased slightly in the early 1980's as the baby-boom generation grew older. Since, as reported in Table 1, the young have relatively high probabilities of leaving unemployment, this accounts for the small movements in $\bar{\pi}$ over this period.

Figure 4 includes a plot of the quarterly averages of monthly values of $\bar{\phi}$. We see that $\bar{\phi}$ follows a general upward trend that is dramatic in the early to mid-1970's, but falls off in the early 1980's. Since the $\bar{\phi}_i$'s are relatively constant over time, equation (20) indicates that this variation in $\bar{\phi}$ must be the result of variation in labor force shares. Our

[21] This equation is estimated using annual data from 1967 to 1983.

[22] The military variable is zero in the years in which there was no draft (January 1970 to December 1983), and equal to the ratio of the number of military personnel to the male population age 16–34 in years in which a draft was present. The draft after 1970 was a lottery draft. Following Barro (1977, 1978), we include this in the "non-draft" period since the incentives to search were considerably decreased by this process of conscription. Monthly data for this variable is available beginning 1953:8.

[23] In a recent study of micro turnover data, Robert Topel (1984) found results similar to those reported here. Namely, that as age and experience in the work force increased, the likelihood of becoming unemployed (ϕ) decreased by the likelihood of leaving unemployment (π) decreased as well.

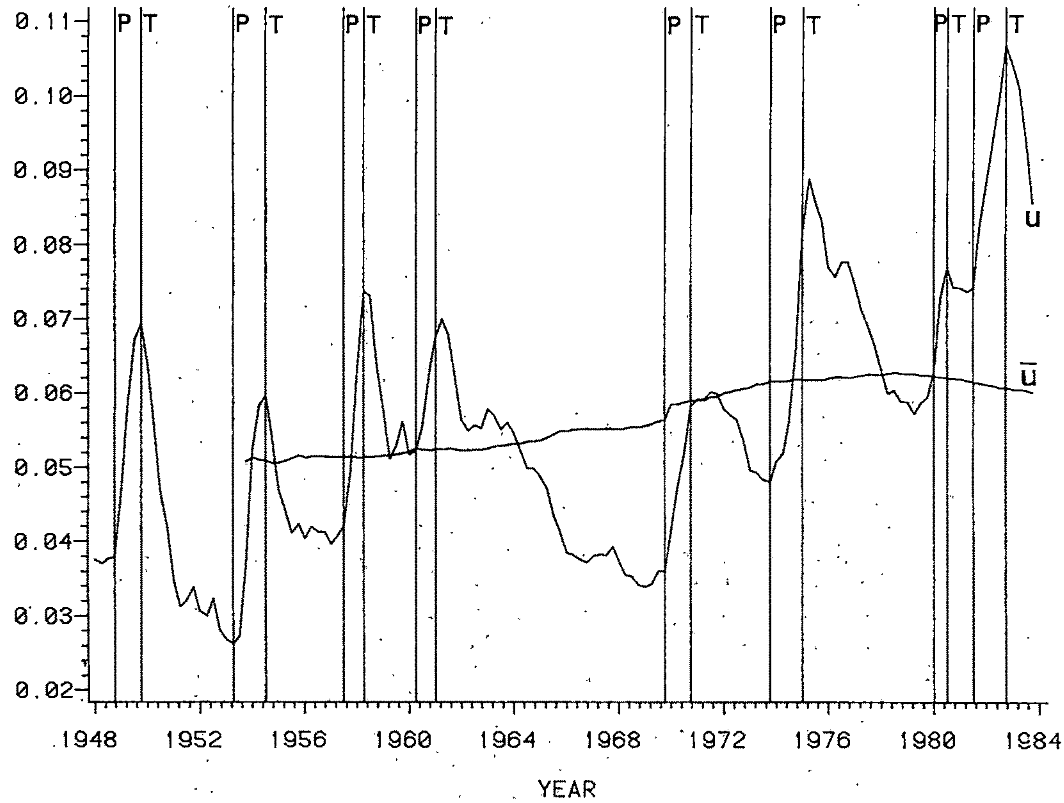


FIGURE 5. ACTUAL AND NATURAL UNEMPLOYMENT RATE

TABLE 1—ESTIMATED NORMAL VALUES OF π AND ϕ BY AGE AND SEX

Age Group	Males				Females			
	$\bar{\gamma}_i$	$\bar{\pi}_i$	$\bar{\phi}_i$	$\bar{u}_i$	$\bar{\gamma}_i$	$\bar{\pi}_i$	$\bar{\phi}_i$	$\bar{u}_i$
16–19[a]	0.0023	0.5366	0.0895	0.1665	0.0028	0.5715	0.0974	0.1701
20–24[a]	0.0036	0.4373	0.0434	0.0988	0.0043	0.5243	0.0506	0.0964
25–34	0.0033	0.3847	0.0179	0.0463	0.0060	0.4932	0.0333	0.0671
35–44	0.0005	0.3571	0.0114	0.0319	0.0024	0.4584	0.0229	0.0498
45–54	−0.0002	0.3379	0.0102	0.0301	0.0012	0.4262	0.0172	0.0403
55–64	0.0003	0.3164	0.0101	0.0320	0.0015	0.3966	0.0136	0.0342
65 and over	−0.0004	0.3449	0.0122	0.0355	0.0016	0.3888	0.0133	0.0342

Note: Calculated as described in text.
[a]Reported values of $\bar{\pi}_i$ and $\bar{\phi}_i$ for young males (16–24) are the values since 1970:1 with *MIL* = 0. Prior to 1970:1, these values varied monthly with the draft variable *MIL*.

examination of the labor force share data reveals that the upswing in $\bar{\phi}$ in the early to mid-1970's is accounted for by the influx of the young and women into the labor force; both are relatively high turnover groups as evidenced by Table 1. Accordingly, the fall-off in $\bar{\phi}$ in the early 1980's is accounted for by the decreased share of the young in the labor force as the baby-boom generation grew older. The estimated ϕ values move more

dramatically than $\bar{\pi}$ because of the much greater variation in ϕ_i than in π_i values.

C. *Evidence of Heterogeneity in* π

The adjustment of the aggregate measure of π to its equilibrium value, $\bar{\pi}$, can be slowed by the presence of heterogeneity in the labor market. As we showed in Section I, Part B, if there are two groups with different values of $\bar{\pi}_i$ and $\bar{\phi}_i$, aggregate measures of π can show persistent deviations from $\bar{\pi}$, even though within groups there is no persistent deviation from equilibrium values. Any attempt then to measure persistence in $\hat{\pi}$, and thus $\hat{u}$, requires an empirical characterization of heterogeneity. In this section we provide empirical evidence of heterogeneity in π.

Let d_v denote the average duration of unemployment this month for those who were unemployed last month and are still unemployed, and let d_{-1} denote the average duration of unemployment in the previous month for all unemployed workers. Suppose that all individuals have the same probability of exiting unemployment, π; that is, there is no heterogeneity. Then,

$$d_v = d_{-1} + 1. \tag{23}$$

Given a uniform escape probability at any point in time, the average duration of unemployment for those remaining unemployed in a given month should increase by one month. However, if individuals are heterogeneous, and thus have different values of π, then d_v will generally exceed $d_{-1}+1$ because a disproportionate share of high-duration (low π) workers remain unemployed at the end of a month's time.

Given a heterogeneous workforce, on average d_v will exceed $(d_{-1}+1)$ by some positive value; call it h. Thus,

$$d_v = d_{-1} + 1 + h. \tag{24}$$

Let d_n denote the average duration of people who became unemployed in the last month. Then the average duration of unemployment in a given month, d, is simply a weighted average of d_v and d_n:

$$d = \frac{s - s^{0-4}}{s}(d_{-1} + 1 + h) + \frac{s^{0-4}}{s} d_n. \tag{25}$$

Solving (25) for h yields

$$h = \frac{d - d_n\left(s^{0-4}/s\right)}{1 - \left(s^{0-4}/s\right)} - d_{-1} - 1. \tag{26}$$

Data on all of the right-hand side variables are readily available, except for d_n which can be approximated and in our sample lies between 0.4267 and 0.4691.[24] To compute the mean value of h over 1953:8 through 1983:12, we assume d_n to be constant at 0.4495, which corresponds to the mean value of π of 0.45674. The resulting mean value, $\bar{h}$, is 0.973. What does this value imply about the extent of heterogeneity in the labor force?

Suppose that the labor force can be characterized as consisting of two groups as in Section I, Part B. It is straightforward to determine the steady-state value of $\bar{h}$ for alternative values of $\bar{\pi}_1$ and $\bar{\pi}_2$, given that the steady-state ratio of inflows $\bar{f}_1/\bar{f}_2$ adjusts so that the aggregate $\bar{\pi}$ is equal to its sample mean of about 0.46.[25] Candidate pairs of values of $\bar{\pi}_1$ and $\bar{\pi}_2$ are those for which the calculated value of $\bar{h}$ is close to 0.973. Table 2 reports these pairs together with the

[24] Those who became unemployed within the month will have an average duration of less than one-half month, because those who became unemployed earlier in the month are more likely to have found a job than the more recently unemployed. Assuming the Poisson distribution of fn. 17 above and $\bar{\gamma} = 0$, the average duration of those who became unemployed within the month is

$$d_n = \frac{\int_0^1 te^{-\theta t}\,dt}{\int_0^1 e^{-\theta t}\,dt} = \frac{1 - e^{-\theta} - \theta e^{-\theta}}{\theta(1 - e^{-\theta})} = \frac{1}{\theta} - \frac{e^{-\theta}}{1 - e^{-\theta}}.$$

Recall that $\theta = -\log(1-\pi)$. The minimum and maximum values of π observed over 1953:8–1983:12 are 0.31 and 0.59 for which the corresponding d_n values are 0.4691 and 0.4267, respectively.

[25] See the Appendix, Part B, for details.

TABLE 2—$\bar{\pi}_1$, $\bar{\pi}_2$, AND $\bar{f}_1/\bar{f}_2$ COMBINATIONS FOR WHICH $\bar{\pi} = 0.45674$ AND $h = 0.973$ IN STATIONARY-STATE EQUILIBRIUM

$\bar{\pi}_1$	$\bar{\pi}_2$	$\bar{f}_1/\bar{f}_2$	$\bar{h}$
0.90	0.225	4.72	0.977
0.85	0.217	4.72	0.976
0.80	0.208	5.01	0.973
0.75	0.197	5.60	0.973
0.70	0.183	6.70	0.976
0.65	0.166	8.70	0.974
0.60	0.143	13.00	0.976
0.55	0.112	24.85	0.968
0.54	0.104	29.96	0.969
0.53	0.095	37.35	0.975
0.52	0.086	47.83	0.971
0.51	0.076	64.52	0.968
0.50	0.064	95.14	0.981
0.49	0.052	153.44	0.975
0.48	0.038	303.88	0.985
0.47	0.023	892.5	0.981
0.46	0.006	14,156.7	0.980

Note: $\bar{\pi}_2$ is the value to three digits which yields $\bar{h}$ nearest to 0.973 given that $\bar{f}_1/\bar{f}_2$ adjusts so that $\bar{\pi} = 0.45675$. See the Appendix, Part B, for details.

implied values of $\bar{f}_1/\bar{f}_2$ and $\bar{h}$. Note that the $\bar{h}$ in our sample implies two groups with substantially different values of $\bar{\pi}_i$.

We can get some idea of potential values for $\bar{\pi}_1$ and $\bar{\pi}_2$ from the range of values reported in Table 1 for broad age-sex cells. These values range from 0.31 to 0.57. Suppose that the average $\bar{\pi}_i$ for 16–19-year olds of both sexes (0.55) reflects $\bar{\pi}_1$, because these youngsters have not yet formed any permanent job attachments nor acquired specific human capital.[26] Table 2 tells us that $\bar{\pi}_2$ would be about 0.112. If the normal unemployment rate among group 2 were about half the overall rate (say 2.75 percent), then $\bar{\phi}_2$ would be 0.00308. Put differently, the average duration of these permanent jobs would be about 325 months or 27 years. This number is consistent with the stylized facts about the labor market.[27] A $\bar{\pi}_2$ of 0.112 would imply that during the recovery from a recession when $\bar{\pi}_2$ dominates the adjustment rate, cyclical unemployment would fall by about one-ninth in one month, one-third in a quarter, one-half in a half-year, and three-quarters in a year's time if all ϕ's and π's were at their normal values. Such persistence may be sufficient to explain the puzzle of why unemployment is persistent without need for recourse to any substantially persistent expectational errors which cause π_1 and π_2 to differ from their normal values for long periods of time. This hypothesis is considered further in Section III.

A final implication of using $\bar{\pi}_1 = .55$ and $\bar{\pi}_2 = .112$ is that normally 79 percent of the unemployed would be from group 1 and only 21 percent from group 2.[28] However, if we assume that the π_i values were at their normal levels when π fell to 0.313 in November 1982, then we conclude that s_1/s fell to only about 0.459. The unemployment rate of 10.7 percent at that time can be decomposed into 4.91 percentage points due to group 1 and 5.79 percentage points due to group 2. Group 1 accounts for 4.77 percentage points of the 6.07 percentage point value of $\bar{u}$, but for only 0.14 percentage points of the total 4.63 percentage points of cyclical unemployment.[29] Group 2 accounts for the remaining 1.30 percentage points of normal unemployment and 4.49 percentage points of cyclical unemployment. This illustrates that the data are consistent with our theoretical analysis in Section I, Parts A and B: normal unemployment consists primarily of high-turnover individuals, but cyclical unemployment is dominated during the recovery from a recession by individuals who have lost permanent jobs and are searching for a replacement. Loss of permanent jobs during recessions will be more prevalent as firms in declining

[26] The correct $\bar{\pi}_1$ could be even a bit lower than this if the average high-probability individual has a lower probability than these youngsters of getting a job or leaving the labor force.

[27] See Robert Hall (1982).

[28] This is computed as $\bar{s}_1/\bar{s} = (0.45674 - 0.112)/(0.55 - 0.112) = 0.787$.

[29] We calculate $s_1/s = (0.313 - 0.112)/(0.55 - 0.112) = 0.459$; $\bar{s}_1/\bar{n} = (\bar{s}_1/\bar{s})(\bar{u}) = (0.787)(6.066\%) = 4.77\%$; $s_1/n = (s_1/s)(u) = (0.459)(10.7\%) = 4.91\%$.

industries find it profitable to accelerate the eventual reductions in their labor force during such times.[30] The search process for such permanent job losers—possibly involving relocation or shifting industries—is a lengthy one so that cyclical unemployment falls much more slowly than would be suggested by a normal $\bar{\pi}$ of almost one-half.

In this section we developed a method to characterize the extent of heterogeneity empirically using data on the duration of unemployment. We then assessed the impact of the heterogeneity in the 1953–82 period. The result is striking—that the observed heterogeneity could very well account for the observed persistence in the aggregate $\hat{\pi}$. In the next section we develop a regression model that allows us to test this hypothesis as well as characterize other cyclical determinants of $\hat{\pi}$, $\hat{\phi}$, and $\hat{\gamma}$.

III. An Empirical Model of Cyclical Unemployment

The model of unemployment rate dynamics we have developed allows us to identify four separate potential sources of persistence in cyclical unemployment, $\hat{u}$. First, persistence could be due to the partial adjustment of u toward $\bar{u}$ with π, ϕ, and γ at their normal levels. Such persistence is pictured in Figure 1. Our estimation of $\bar{\pi}$ in Section II implies that, on average, 46 percent of the unemployed escape unemployment each month which would imply that, during recovery, nearly half the cyclically unemployed would find jobs within a month. However, as we demonstrated empirically in Section II, the partial-adjustment speed may be substantially reduced due to heterogeneity, with individual π, ϕ, and γ at normal levels. We illustrated this phenomenon in Figure 2, and the estimates we generated in Section II, Part C, show that only about one-ninth of the cyclically employed would find jobs within a month. A third source of persistence might be autocorrelation in $\hat{\pi}$, $\hat{\phi}$, and $\hat{\gamma}$ due to equilibrium adjustments to changes in economic conditions. Finally, autocorrelation in $\hat{\pi}$, $\hat{\phi}$, and $\hat{\gamma}$ may be associated with persistent expectational errors. The first three sources of persistence are consistent with a rational expectations model of economic agents' behavior, while the fourth contradicts such a hypothesis.

In this section, we complement the measures of $\bar{\pi}$, $\bar{\phi}$, and $\bar{\gamma}$ developed in Section II with equations that explain the movement of π, ϕ, and γ around those normal levels, and with a supplementary equation which links the movements of inventories and money. This preliminary analysis of the data raises many questions for future research, but several messages are clear. First, heterogeneity is an important source of persistence, which confirms the empirical analysis just presented. Second, the data appear to be consistent with equilibrium models of persistence, with little if any role for autocorrelated expectational errors. Finally, within the simple model we have chosen to describe, deviations of the unemployment rate from its natural level are not caused directly by money shocks, but instead through fluctuations in the level of inventories about its mean. Thus aggregate shocks that cause the level of inventories to fluctuate, such as unanticipated fluctuations in the money supply, will affect the cyclical unemployment rate.

A. *Behavior of the Employment Probability $\hat{\pi}$*

The standard search model of unemployment states that $\hat{\pi}$ is a function of unexpected changes in aggregate demand and hence the derived demand for labor. An unexpected increase (decrease) in aggregate demand shifts the actual distribution of wage

[30] This is because a reduction in the current value of the marginal product will have a disproportionately greater effect on the present discounted value of the marginal product for declining industries. Essentially, we are arguing that structural change in the economy will be greater during recessions. The basic idea is that the necessary reallocation of labor associated with changing tastes and technology is likely to be bunched together during recessions. This is because during these periods the value of production is relatively low, and therefore this becomes an optimal time to make changes that were eventually going to be made anyway.

offers right (left) relative to the expected distribution so that π is increased (decreased) relative to normal. Another, possibly complementary, theory posits that the probability of receiving any offer increases (decreases) so that the probability π of an acceptable offer increases (decreases). Following Barro (1977, 1978), we use unexpected money $\tilde{M}$ as an indicator of shifts in aggregate demand. The strictest interpretation of rational expectations would posit that only the current value of $\tilde{M}$ should affect $\hat{\pi}$.[31] We test this hypothesis by including in our $\hat{\pi}$ regression specification a distributed lag on the current value and first 11 lags of $\tilde{M}$.[32]

The cyclical component $\hat{I}$ of the inventory-sales ratio may affect $\hat{\pi}$ for three reasons.[33] First, our measure of π is actually a weighted average of the probability that a searcher finds an acceptable job, the probability that a searcher leaves the labor force, and the probability that a laid-off worker is recalled. The probability of recall from layoff depends positively on the extent to which excess inventories have been eliminated.[34] Second, when inventories are abnormally high, the discounted marginal value product of labor is abnormally low. Optimal intertemporal leisure substitution implies that some searchers should then drop out of the labor force, but we expect that in fact many of these individuals would be counted as unemployed due to unemployment benefit rules. Finally, cyclical fluctuations in inventories could shift the actual wage offer distribution relative to the expected distribution.

To capture these inventory effects, the $\hat{\pi}$ equation also includes a 12-month distributed lag on $\hat{I}$. Intertemporal substitution considerations would suggest that $\hat{I}$ should enter because it is the most recent information available to workers during the month. From the point of view of firms, lagged as well as current values of $\hat{I}$ could enter because, if there were adjustment costs associated with varying employment, it would be optimal for firms to use inventories to buffer short-run monthly variations in sales. However, given that there are also costs of varying inventories, successive monthly variations in the inventory-sales ratio in the same

[31] For such an interpretation, see Bennett McCallum (1979). The basic idea is that all past information will be incorporated in current expectations and so will have no effect on real variables except through past effects on current state variables such as inventories. If lagged values of $\tilde{M}$ were to enter, we would conclude (a) expectations are not formed rationally, (b) the relevant horizon for forming expectations is longer than 1 month, or (c) some significant state variables have been omitted from the regression. In the latter case, the lagged $\tilde{M}$ coefficients would reflect past effects on the omitted state variable(s) and current effects of the state variable(s).

[32] We measure $\tilde{M}$ as the residual from an ARIMA(0,2,4) process fit to log M. Our money series is the current Federal Reserve M_1 series for 1959:1 through 1983:12 which we have extended back to 1953:6 by a ratio splice at 1959:1 to the old M_1 series in Board of Governors of the Federal Reserve System (1976). (This splice preserves the growth rates in the two series which are practically identical in 1959.) After allowing for taking second differences, $\tilde{M}$ estimates are available from 1953:8. This start date for estimating $\tilde{M}$ was chosen for two reasons: 1) this avoids essentially all of the period during which the Fed was pegging government bond prices. The process determining money growth is potentially different during that period; see Milton Friedman and Anna Schwartz (1963, pp. 613, 625). 2) This is the latest start date for $\tilde{M}$ for which availability of $\tilde{M}$ observations does not reduce the period over which we can estimate the regressions reported in this section.

[33] We measure $\hat{I}$ as the deviation from a linear trend fit to Citibase data on the inventory-sales ratio for manufacturing and trade. This ratio (the one-month lagged value of *Business Conditions Digest* series number 77) has the beginning-of-month total book value (in 1972 dollars) of manufacturing and trade inventories as the numerator and manufacturing and trade sales for the prior month (in 1972 dollars) as the denominator. This dating is appropriate both with respect to the decision making of the firm and for the intertemporal decision making of individuals. The linear trend was not estimated directly, but instead is that $\bar{I}$ implied by our regression of I on both trend and cyclical variables (see Table 5 below). The difference is not a substantive one: the correlation coefficient between our $\hat{I}$ and the residuals from a linear trend regression is 0.95. Details of how to impute $\bar{I}$ from the Table 5 regression are given in fn. 41 below.

[34] See Haltiwanger and Louis Maccini (1984) and Topel (1982) for theoretical results which support the discussion of the interaction of inventories and layoffs in this section.

TABLE 3—DETERMINANTS OF $\hat{\pi}$

	With			With	
	$\hat{I}$ and $\tilde{M}$ (1)	$\hat{I}$ Only (2)		$\hat{I}$ and $\tilde{M}$ (1)	$\hat{I}$ Only (2)
Constant	−0.4588	−0.4582	$\tilde{M}$	0.5047	
	(0.0112)	(0.0110)		(0.3727)	
Lagged Share of Unemployed with 0–4 Weeks Duration	1.0073	1.0066	$\tilde{M}(-1)$	0.3039	
	(0.0246)	(0.0242)		(0.3739)	
$\hat{I}$	−0.5677	−0.6260	$\tilde{M}(-2)$	−0.3642	
	(0.1296)	(0.1220)		(0.3747)	
$\hat{I}(-1)$	0.1152	0.1557	$\tilde{M}(-3)$	−0.1325	
	(0.1642)	(0.1606)		(0.3758)	
$\hat{I}(-2)$	−0.0870	−0.0481	$\tilde{M}(-4)$	0.3922	
	(0.1646)	(0.1605)		(0.3793)	
$\hat{I}(-3)$	0.1409	0.1264	$\tilde{M}(-5)$	0.0711	
	(0.1648)	(0.1606)		(0.3885)	
$\hat{I}(-4)$	0.1287	0.0933	$\tilde{M}(-6)$	0.1689	
	(0.1653)	(0.1610)		(0.3929)	
$\hat{I}(-5)$	−0.1099	−0.1013	$\tilde{M}(-7)$	−0.0402	
	(0.1654)	(0.1608)		(0.3962)	
$\hat{I}(-6)$	0.2245	0.2135	$\tilde{M}(-8)$	−0.2048	
	(0.1664)	(0.1624)		(0.3932)	
$\hat{I}(-7)$	−0.1341	−0.1044	$\tilde{M}(-9)$	0.2350	
	(0.1656)	(0.1616)		(0.3974)	
$\hat{I}(-8)$	0.0001	0.0058	$\tilde{M}(-10)$	−0.0798	
	(0.1670)	(0.1617)		(0.3962)	
$\hat{I}(-9)$	0.1500	0.1157	$\tilde{M}(-11)$	0.4122	
	(0.1675)	(0.1610)		(0.1296)	
$\hat{I}(-10)$	−0.0935	−0.0390	$\bar{R}^2$	0.850	0.852
	(0.1683)	(0.1614)	$S.E.E.$	0.026	0.026
$\hat{I}(-11)$	0.1630	0.1324	D-W	1.828	1.833
	(0.1255)	(0.1193)			

Note: Standard errors are shown in parentheses. Period of estimation is 1954:8 through 1983:12.

direction will make it optimal for a firm eventually to change production and hence employment. To capture this optimal lagged response of employment to inventories, we include a distributed lag of $\hat{I}$ as determinants of $\hat{\pi}$.

To capture the effects of heterogeneity on $\hat{\pi}$, we require variables that indicate how the composition of the unemployed has changed over time. In particular, we are interested in changes in the share of the unemployed who exhibit high turnover and low duration relative to the share of the unemployed who exhibit low turnover and high duration. To characterize this simply, we include the one-period lagged share of unemployed who have been unemployed less than 5 weeks as an explanatory variable in the $\hat{\pi}$ equation.

Column 1 of Table 3 reports the results of *OLS* regressions of $\hat{\pi}$ on the lagged unemployment share of low duration workers, and the 12-month distributed lags on $\hat{I}$ and $\tilde{M}$ using *OLS*. The coefficients on the unemployment share is positive indicating that $\hat{\pi}$ increases as the share of the short duration unemployed increases. The coefficient is significant at the 99 percent confidence level. Neither the current money shock nor all 12 coefficients as a group are significantly different from zero.[35] The coefficient on $\hat{I}$ is negative and significant and the distributed lag

[35]The $F(12,327)$-statistic for testing the null hypothesis that all the coefficients on $\tilde{M}$ are zero is 0.570.

on $\hat{I}$ is significant as well.[36] In column (2) we report a regression which omits the insignificant lags on money shocks. The $\hat{I}$ distributed lag is now significant at the 1 percent confidence level.[37] Inclusion of additional lagged unemployment shares by other duration intervals does not alter the qualitative results reported in Table 3 (the insignificant difference from 1 of the $(s^{0-4}/s)_{-1}$ coefficient neither is nor should be robust across specifications).

We interpret these results as indicating that the primary factor determining cyclical variations in the probability of leaving unemployment is heterogeneity. Inventory innovations appear to play some role and surprisingly, money shocks have no significant impact. Money shocks may be insignificant because they operate only through the $\hat{I}$, or because of measurement error problems, but there is certainly no evidence here of persistent expectational errors.

B. *Behavior of* $\hat{\phi}$

Since firings and layoffs are the complement of firms' decisions with respect to new hires and recalls, we would expect $\hat{\phi}$ to be increased by high cyclical inventories and decreased by positive money shocks. High inventories or low sales due to a negative $\tilde{M}$ present the best time for firms to cull their labor force of marginal workers; when these conditions are persistent, layoffs (permanent and temporary) will result. In the opposite direction, low $\hat{I}$ and high $\tilde{M}$ would tend to induce firms to retain otherwise unsatisfactory workers temporarily and would reduce the aggregate incidence of new layoffs below its normal level. So as with $\hat{\pi}$, we include 12-month distributed lags on $\hat{I}$ and $\tilde{M}$ in our regressions explaining $\hat{\phi}$.

Heterogeneity in the labor force in terms of turnover propensities is a third factor which may influence $\hat{\phi}$. Our measure of $\hat{\phi}$ already controls for such heterogeneity that is associated with the age-sex composition of the labor force. However, other characteristics of individuals may be related to heterogeneity in turnover propensities. In an attempt to capture some of this residual heterogeneity, we include measures of the industrial composition of the labor force in the $\hat{\phi}$ regressions.

Column 1 of Table 4 reports the results of regressing $\hat{\phi}$ on the 12-month distributed lags on $\hat{I}$ and $\tilde{M}$, and on the industrial composition of the labor force. The shares of the labor force by industry proved to be significant indicating heterogeneity in turnover propensities across industries.[38] Relatively high-turnover industries include construction, mining, and manufacturing (nondurables), whereas relatively low-turnover industries include transportation and manufacturing (durables). As a group, the distributed lag on $\hat{I}$ proved quite significant in explaining $\hat{\phi}$.[39] The large, significant, positive coefficients on $\hat{I}$ and $\hat{I}(-1)$ reflect the increase in layoffs associated with high cyclical inventories. It is not possible to decompose this into temporary and permanent layoffs, but the significant negative coefficient on $\hat{I}(-11)$ suggests that variation in the rate of permanent separations for marginal workers is involved: if a marginal worker is fired earlier due to high cyclical inventories, $\hat{\phi}$ will decrease below normal levels during the latter period when the worker would have otherwise been fired. The sum of the coefficients is positive confirming the notion that high cyclical inventories cause temporary layoffs of employees who otherwise would never have been unemployed. A similar pattern of coefficients appears in the distributed lag on $\tilde{M}$. A high value of $\tilde{M}$ reduces the likelihood that workers will be terminated in the first few months, but there

[36]The $F(12,327)$-statistic for testing the null hypothesis that all of the coefficients on $\hat{I}$ are zero is 2.974.

[37]The $F(12,339)$-statistic for testing the null hypothesis that all of the coefficients on $\bar{I}$ are zero is 9.771.

[38]The $F(7,321)$-statistic for testing the null hypothesis that all labor force share by industry coefficients are zero is 161.39.

[39]The $F(12,321)$-statistic for testing the null hypothesis that all $\hat{I}$ coefficients are zero is 16.01.

TABLE 4—DETERMINANTS OF $\hat{\phi}$

	With $\hat{I}$ and $\tilde{M}$ (1)	$\hat{I}$ Only (2)		With $\hat{I}$ and $\tilde{M}$ (1)	$\hat{I}$ Only (2)
Constant	−0.0325	−0.0277	$\tilde{M}$	−0.0012	–
	(0.04363)	(0.0425)		(0.0209)	
LF Share of Transportation	−0.5475	−0.5295	$\tilde{M}(-1)$	−0.0155	–
	(0.0966)	(0.0946)		(0.0210)	
LF Share of Finance and Services	0.1107	0.1056	$\tilde{M}(-2)$	−0.0139	–
	(0.0484)	(0.0468)		(0.0210)	
LF Share of Manufacturing-Durables	−0.3020	−0.3052	$\tilde{M}(-3)$	−0.0420	–
	(0.0669)	(0.0644)		(0.0213)	
LF Share of Construction	0.3513	0.3474	$\tilde{M}(-4)$	−0.0109	–
	(0.0762)	(0.0746)		(0.0215)	
LF Shares of Mining	1.1473	1.1457	$\tilde{M}(-5)$	−0.0275	–
	(0.1324)	(0.1298)		(0.0221)	
LF Share of Wholesale and Retail Trade	0.0171	0.0062	$\tilde{M}(-6)$	−0.0102	–
	(0.0850)	(0.0842)		(0.0225)	
LF Shares of Manufacturing-Nondurables	0.8229	0.8054	$\tilde{M}(-7)$	0.0147	–
	(0.0590)	(0.0579)		(0.0227)	
$\hat{I}$	0.0355	0.0376	$\tilde{M}(-8)$	0.0165	–
	(0.0072)	(0.0067)		(0.0223)	
$\hat{I}(-1)$	0.0187	0.0196	$\tilde{M}(-9)$	−0.0025	–
	(0.0092)	(0.0090)		(0.0226)	
$\hat{I}(-2)$	−0.0062	−0.0077	$\tilde{M}(-10)$	0.0047	–
	(0.0091)	(0.0089)		(0.0226)	
$\hat{I}(-3)$	−0.0039	−0.0038	$\tilde{M}(-11)$	0.0217	–
	(0.0091)	(0.0089)		(0.0223)	
$\hat{I}(-4)$	0.0044	−0.0014			
	(0.0092)	(0.0090)	Sum of *DL* Coefficients:		
			on $\hat{I}$	0.0310	0.0303
$\hat{I}(-5)$	−0.0078	−0.0074	on $\tilde{M}$	−0.0442	–
	(0.0092)	(0.0090)			
$\hat{I}(-6)$	0.0032	−0.0024	$\bar{R}^2$	0.790	0.792
	(0.0093)	(0.0091)			
$\hat{I}(-7)$	−0.0055	−0.0071	*S.E.E.*	0.001	0.001
	(0.0092)	(0.0090)			
$\hat{I}(-8)$	0.0065	0.0079	*D-W*	1.217	1.249
	(0.0093)	(0.0090)			
$\hat{I}(-9)$	0.0013	0.0023			
	(0.0093)	(0.0090)			
$\hat{I}(-10)$	−0.0081	−0.0061			
	(0.0094)	(0.0090)			
$\hat{I}(-11)$	−0.0233	−0.0210			
	(0.0072)	(0.0068)			

Note: Standard errors are shown in parentheses. Period of estimation is 1954:8 through 1983:12.

is essentially no effect over a year's time as indicated by the insignificance of money shocks as a group.[40] Column 2 of Table 4 reports the results of a regression which drops the insignificant $\tilde{M}$ distributed lag. The S-shaped pattern of $\hat{I}$ coefficients is more pronounced than in the previous regression. Again we conclude that lagged expectational errors from money shocks account for no detectible persistence in the unemployment rate except as it is incorporated in $\hat{I}$.

[40]The $F(12,321)$-statistic for testing the null hypothesis that all $\tilde{M}$ coefficients are zero is 0.716.

C. *Cyclical Behavior of* $\hat{\gamma}$

Given potential discouraged and added worker effects over the cycle, both $\hat{I}$ and $\tilde{M}$ might influence $\hat{\gamma}$. In regressions not reported here, we tried regressing $\hat{\gamma}$ on 12-month distributed lags on these variables. The *F*-statistic on each group of parameters was insignificant and the $\bar{R}^2$ negligible. We conclude that, while various subgroups of the labor force may have cyclically sensitive participation rates, the overall growth rate of the labor force is not sensitive to cyclical factors.

D. *Cyclical Behavior of the Inventory-Sales Ratio* $\hat{I}$

Our results so far indicate that to understand the cyclical behavior of the unemployment rate, the determinants of the inventory-sales ratio must be understood as well. Much research on the cyclical behavior of I is already underway; so, for our present purposes, it suffices to estimate a simple partial-adjustment regression of I on time, the lagged dependent variable, and a 12-month distributed lag on $\tilde{M}_{-1}$ as our indicator of aggregate demand shocks. Since I is the ratio of beginning inventories to last month's sales, the current money shock $\tilde{M}$ occurs too late to affect its value. This regression is reported in Table 5.

TABLE 5—DETERMINANTS OF I

Constant	0.0340
	(0.0072)
Time	0.00004
	(0.00009)
I_{-1}	0.9086
	(0.0190)
$\tilde{M}$: $\tilde{M}(-1)$	−0.4129
	(0.1583)
$\tilde{M}(-2)$	−0.3672
	(0.1577)
$\tilde{M}(-3)$	−0.3467
	(0.1563)
$\tilde{M}(-4)$	−0.4125
	(0.1559)
$\tilde{M}(-5)$	−0.5682
	(0.1591)
$\tilde{M}(-6)$	−0.4715
	(0.1605)
$\tilde{M}(-7)$	−0.3466
	(0.1622)
$\tilde{M}(-8)$	−0.2325
	(0.1657)
$\tilde{M}(-9)$	−0.3219
	(0.1655)
$\tilde{M}(-10)$	−0.0832
	(0.1664)
$\tilde{M}(-11)$	−0.1791
	(0.1686)
$\tilde{M}(-12)$	−0.2422
	(0.1695)
Sum of *DL* Coefficients on $\tilde{M}$:	−3.985
$\bar{R}^2$	0.950
S.E.E.	0.012
Durbin's *h*	−1.569

Note: Standard errors are shown in parentheses. Period of estimation is 1954:8 through 1983:12.

Although the equation is a simple one, it in fact explains the behavior of the inventory-sales ratio well. The coefficient on I_{-1} indicates that 9 percent of cyclical inventories are eliminated per month; this corresponds to 25 percent in three months, 44 percent in six months, and 68 percent in a year's time. The long-run effect of the time trend term is only 0.0004 per annum.[41] The distributed lag on $\tilde{M}$ indicates that positive money shocks significantly decrease inventories from the first through seventh month. The gradual build up of production relative to final sales is consistent with shock-absorber money demand and costs of changing production levels.[42]

[41] If we define $\bar{I}_t$ as the value to which I_t converges in the absence of any money shocks or random disturbances, we can write $\bar{I}_t = \alpha + \beta t$ in this case. If in the specified regression equation, a is the constant, b the coefficient of time, and c the coefficient of I_{-1}, it can be shown that the long-run values are found as $\beta = b/(1-c)$ and $\alpha = (a/(1-c)) - (bc/(1-c)^2)$. The $bc/(1-c)^2$ adjustment in computing α arises because normal growth in $\bar{I}$ is conventionally included in the constant term instead of appearing explicitly in the partial-adjustment mechanism. (Were it included there, we would have a partial-adjustment term like $(1-c)(\bar{I}_t - \beta - I_{t-1})$.) Our estimates imply $\alpha = 0.3682$ and $\beta = 0.0004$.

[42] On the shock-absorber approach to money demand, see Darby (1972) and Jack Carr and Darby (1981).

IV. Conclusions

This paper has demonstrated that the fraction ϕ of the labor force becoming unemployed in a month and the probability π of leaving unemployment over a month are useful constructs for understanding how the unemployment rate fluctuates around its normal level. Furthermore, these variables have empirical counterparts which we can readily calculate from the aggregate data on numbers and duration of unemployment.

In addition, we have shown that significant heterogeneity across individuals in π implies a much slower convergence of u toward $\bar{u}$ than would be hypothesized from current or normal values of π. We demonstrate that this heterogeneity is quite substantial in the aggregate data.

The hypotheses supported by our initial exploration of these new data can be summarized by means of Figure 6. In this figure we distinguish between two groups: the first is characterized by high values of $\bar{\phi}_1$ and $\bar{\pi}_1$ compared to the second group. Money shocks do not directly affect the proximate contemporaneous determinants ($\hat{\pi}_i$ and $\hat{\phi}_i$) of each group's cyclical unemployment rate $\hat{u}_i$, but do so indirectly through cyclical inventories $\hat{I}$. A restrictive monetary policy causes $\hat{I}$ to gradually build up. High $\hat{I}$ values may depress the $\hat{\pi}_i$ directly, but their main effect seems to be on the cyclical search rates $\hat{\phi}_1$ and $\hat{\phi}_2$. High values of the $\hat{\phi}_i$ (and perhaps negative $\hat{\pi}_i$) build up $\hat{u}_1$ and especially $\hat{u}_2$ for a period of some months. Then as the effect of the money shock on inventories is attentuated, u_1 quickly returns to normal ($\hat{u}_1$ goes to 0). But the low-turnover group is characterized by a very low normal probability $\bar{\pi}_2$, and this value governs the speed of adjustment of $\hat{u}_2$ (and eventually $\hat{u}$) toward 0.

Indeed, it appears that in normal times the bulk of unemployment is comprised of group 1 individuals. Major recessions have the effect of disemploying very large numbers of group 2 individuals who otherwise have nearly permanent jobs. Their lengthy process of search for a new permanent job appears to dominate the recovery period and explain substantial persistence in unemployment. It is not that any individual is taking unusually long to find a job; it is simply unusual to have so many slow searchers unemployed at once.

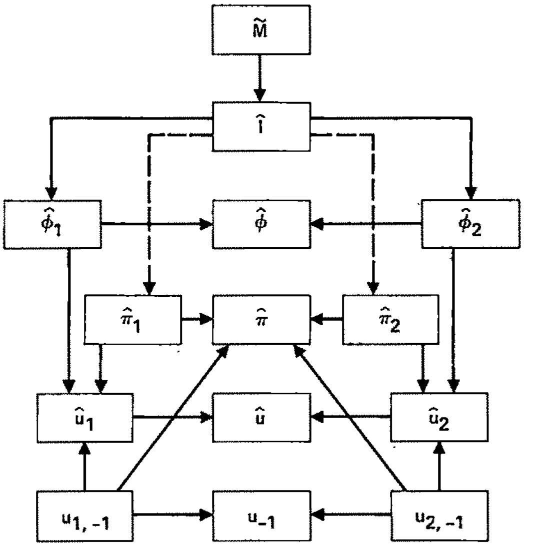


FIGURE 6. DETERMINATION OF GROUP AND AGGREGATE VALUES OF ϕ, π, u

To return to Figure 6, we note that given exogenous labor force shares and growth rates, we can infer the aggregate values of $\hat{\phi}$, $\hat{\pi}$, and $\hat{u}$ from the corresponding values for each group. At present, we are constrained to work with these aggregate data and infer compositional effects through proxies and other indirect evidence. Our first task for future research will be to develop new measures and evidence to permit us to observe more directly the nature of labor force heterogeneity and its influence on unemployment rate dynamics.

For now we conclude that the observed persistence in unemployment appears to be consistent with equilibrium models and rational expectations since we are unable to detect any effect of lagged—or even current—money shocks on $\hat{\pi}$. Inventories appear to be the key channel transmitting the effects of money shocks to the proximate determinants of unemployment.

Appendix

A. *Calendar Bias in Measured π Due to Survey Timing*

The potential calendar bias in π is associated with our estimate of s_{-1}, which is supposed to be the number unemployed exactly one standard month ago (4.35 weeks). However, in months in which there are 4(5) weeks between surveys, our estimates of s_{-1} is actually the number unemployed 4(5) weeks ago. Denote u_{-1}, u^4_{-1}, and u^5_{-1} as the unemployment rate 4.35, 4, and 5 weeks ago, respectively, and π^4 and π^5 as our estimates of π for 4-week and 5-week intervals between surveys, respectively. Then, neglecting any growth in the labor force within a week the calendar bias is given by

$$\pi - \pi^4 = (1-\pi)\big((u_{-1}/u^4_{-1})-1\big),$$

$$\text{or} \quad \pi - \pi^5 = (1-\pi)\big((u_{-1}/u^5_{-1})-1\big).$$

Note first that there is no calendar bias in the stationary state. Moreover, if, as we believe is generally the case, $(1-\pi)$ is approximately 0.5 and either $((u_{-1}/u^4_1)-1)$ or $((u_{-1}/u^5_{-1})-1)$ is less than 0.04 in absolute value, then we can conclude that the calendar bias is negligible in magnitude.[43]

B. *Calculations of Heterogeneity Statistics*

We measure heterogeneity across individual π_i values by h, the amount by which the unemployment duration d_v of those currently unemployed since last month, exceeds one month plus the average unemployment duration of all persons last month (d_{-1}). This h, which is 0 if all π_i values are equal, has a statistical interpretation in terms of the covariance between duration and changes in shares by duration which can be useful for certain problems.[44] However, we need not be concerned with that interpretation to understand the derivation of Table 2 in the text.

The table is derived by considering the steady-state solution to the two-group heterogeneity model of Section I, Part B. Define $\bar{\theta}_1$ and $\bar{\theta}_2$ to be the continuous time equivalents of $\bar{\pi}_1$ and $\bar{\pi}_2$, respectively. Hence $\bar{\theta}_1 = -\log(1-\bar{\pi}_1)$ and $\bar{\theta}_2 = -\log(1-\bar{\pi}_2)$. Given $\bar{\theta}_1$ and $\bar{\theta}_2$ we can write $\bar{\pi}$ as

$$\text{(A1)} \quad \bar{\pi} = \frac{(\bar{f}_1/\bar{f}_2)(\bar{\pi}_1/\bar{\theta}_1)+(\bar{\pi}_2/\bar{\theta}_2)}{(\bar{f}_1/\bar{f}_2)/\bar{\theta}_1+1/\bar{\theta}_2},$$

which in turn allows us to write $(\bar{f}_1/\bar{f}_2)$ as

$$\text{(A2)} \quad (\bar{f}_1/\bar{f}_2) = (\bar{\theta}_1/\bar{\theta}_2)[(\bar{\pi}_2-\bar{\pi})/(\bar{\pi}-\bar{\pi}_1)]$$

Hence, (A2) defines the steady-state ratio of inflows $\bar{f}_1/\bar{f}_2$ that is consistent with given values of $\bar{\pi}_1$, $\bar{\pi}_2$, and $\bar{\pi}$.

Define $\bar{d}^1$ and $\bar{d}^2$ to be the steady-state average duration of group 1 and 2, respectively. Then $\bar{d}^1 = 1/\bar{\theta}_1$ and $\bar{d}^2 = 1/\bar{\theta}_2$. Also, since all individuals within each group are assumed to be homogeneous, then $\bar{d}^1_v = \bar{d}^1+1$ and $\bar{d}^2_v = \bar{d}^2+1$. Define $\bar{\sigma}^1$ as the steady-state share of unemployed from group 1. Then,

$$\text{(A3)} \quad \bar{\sigma}^1 = \frac{(\bar{f}_1/\bar{f}_2)/\bar{\theta}_1}{(\bar{f}_1/\bar{f}_2)/\bar{\theta}_1+1/\bar{\theta}_2},$$

and $\bar{\sigma}^2 = 1-\bar{\sigma}^1$. This allows us to write d as

$$\text{(A4)} \quad d = \sigma^1 d^1 + \sigma^2 d^2$$

$$= \frac{(\bar{f}_1/\bar{f}_2)/(\bar{\theta}_1)^2+(1/\bar{\theta}_2)^2}{(\bar{f}_1/\bar{f}_2)/\bar{\theta}_1+1/\bar{\theta}_2}.$$

[43] An increase of 1 percent (.01) per day translates into a 30 percent increase in the unemployment rate over a month. This would be 1.5 percentage points on a base of 5 percentage points.

[44] Indeed, it can be shown that $h = d_v - d_{-1} - 1 = m\,\mathrm{cov}(d_{i,-1}, \Delta\sigma_i)$, where $d_{i,-1}$ refers to one of the m durations observed last month and $\Delta\sigma_i$ refers to (a) the ratio of people with duration $d_{i,-1}+1$ this month to total people with duration of one month or over, minus (b) the ratio of number unemployed last month with duration $d_{i,-1}$ to the total unemployed last month.

Given (A2) and the definitions of $\bar{\theta}_1$ and $\bar{\theta}_2$, (A4) allows us to compute the steady-state value of d associated with any combination of $\bar{\pi}_1$, $\bar{\pi}_2$, and $\bar{\pi}$.

In a similar fashion we can define $\bar{\sigma}_v^1$ and $\bar{\sigma}_v^2$ as the share of last period's unemployed from groups 1 and 2, respectively. This, in turn, allows us to write d_v as

$$\text{(A5)}\quad d_v = \bar{\sigma}_v^1 \bar{d}_v^1 + \bar{\sigma}_v^2 \bar{d}_v^2$$
$$= \left\{\left[(1+\bar{\theta}_1)(1-\bar{\pi}_1)(\bar{f}_1/\bar{f}_2)/(\bar{\theta}_1)^2\right] + \left[(1+\bar{\theta}_2)(1-\bar{\pi}_2)/(\bar{\theta}_2)^2\right]\right\}$$
$$/\left\{\left[(1-\bar{\pi}_1)(\bar{f}_1/\bar{f}_2)/\bar{\theta}_1\right] + \left[(1-\bar{\pi}_2)/\bar{\theta}_2\right]\right\}.$$

Since $\bar{h} = \bar{d}_v - \bar{d} - 1$, using (A2), (A4), and (A5), we can calculate $(\bar{f}_1/\bar{f}_2)$ and $\bar{h}$ for any given combination of $\bar{\pi}_1$, $\bar{\pi}_2$, and $\bar{\pi}$. In Table 2 we use the estimated value of $\bar{\pi} = 0.45674$ and $(\bar{f}_1/\bar{f}_2)$ and $\bar{h}$ are calculated in this manner for the values of $\bar{\pi}_1$ and $\bar{\pi}_2$ as given.

REFERENCES

Akerlof, George and Main, Brian G. M., "Unemployment Spells and Unemployment Experience," *American Economic Review*, December 1980, *70*, 885–93.

Barro, Robert J., "Unanticipated Money Growth and Unemployment in the United States," *American Economic Review*, March 1977, *67*, 101–15.

______, "Unanticipated Money, Output and the Price Level in the United States," *Journal of Political Economy*, August 1978, *86*, 549–80.

Barron, John M., "Search in the Labor Market and the Duration of the Unemployment: Some Empirical Evidence," *American Economic Review*, December 1975, *65*, 934–42.

Blanchard, Olivier, "The Production and Inventory Behavior of the American Automobile Industry," *Journal of Political Economy*, June 1983, *91*, 365–400.

Carlson, John A. and Horrigan, Michael W., "Measures of Unemployment Duration as Guides to Research and Policy: Comment," *American Economic Review*, December 1983, *73*, 1143–50.

Carr, Jack and Darby, Michael R., "The Role of Money Supply Shocks in the Short-Run Demand for Money," *Journal of Monetary Economics*, September 1981, *8*, 183–99.

Clark, Kim B. and Summers, Lawrence H., "Labor Market Dynamics and Unemployment: A Reconsideration," *Brookings Papers on Economic Activity*, 1:1979, 13–60.

Darby, Michael R., "The Allocation of Transitory Income Among Consumers' Assets," *American Economic Review*, December 1972, *62*, 928–41.

______, **Lothian, James R., et al.,** *The International Transmission of Inflation*, Chicago: University of Chicago Press, 1983.

Diamond, Peter A., "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, October 1982, *90*, 881–94.

Feldstein, Martin, "Temporary Layoffs in the Theory of Unemployment," *Journal of Political Economy*, October 1976, *84*, 937–57.

Friedman, Milton and Schwartz, Anna J., *A Monetary History of the United States, 1867–1960*, NBER Studies in Business Cycles, Vol. 12, Princeton: Princeton University Press, 1963.

Hall, Robert E., "The Importance of Lifetime Jobs in the U.S. Economy," *American Economic Review*, September 1982, *72*, 724–25.

Haltiwanger, John and Maccini, Louis, "A Model of Inventory and Layoff Behavior under Uncertainty," Working Paper No. 321, University of California-Los Angeles, January 1984.

______ **and Plant, Mark,** "How Should We Measure Slackness in the Labor Market?," Working Paper No. 343, University of California-Los Angeles, September 1984.

Heckman, James J. and Borjas, George J., "Does Unemployment Cause Future Unemployment? Definitions, Questions, and Answers from a Continuous Time Model of Heterogeneity and State Dependence," *Economica*, August 1980, *47*, 247–83.

Howitt, Peter, "Transaction Costs in the Theory of Unemployment," *American Eco-*

nomic Review, March 1985, *75*, 88–100.

Leijonhufvud, Axel, *Information and Coordination*, New York: Oxford University Press, 1981.

Lilien, David M., "Sectoral Shifts and Cyclical Unemployment," *Journal of Political Economy*, August 1982, *90*, 777–93.

Lippman, Steven A. and McCall, John J., *The Economics of Search*, Cambridge: Harvard University Press, 1985.

______ and ______, "Job Search in a Dynamic Economy," *Journal of Economic Theory*, June 1976, *12*, 365–90.

McCallum, Bennett T., "On the Observational Inequivalence of Classical and Keynesian Models," *Journal of Political Economy*, April 1979, *87*, 395–402.

Ries, John C., "Unemployment in 1982: Beyond the Official Labor Force Statistics," *New England Economic Review*, May/June 1984, 29–37.

Salant, Stephen, "Search Theory and Duration Data: A Theory of Sorts," *Quarterly Journal of Economics*, February 1977, *91*, 39–58.

Smith, R. and Vanski, J., "Gross Change Data: The Neglected Data Base," in National Commission on Employment and Unemployment Statistics, *Counting the Labor Force*, Vol. II, Washington: USGPO, 1979.

Topel, Robert, "Inventories, Layoffs and the Short-Run Demand for Labor," *American Economic Review*, September 1982, *72*, 769–87.

______, "Equilibrium Earnings, Turnover, and Unemployment: New Evidence," *Journal of Labor Economics*, October 1984, *2*, 500–22.

Wachter, Michael L., "The Changing Cyclical Responsiveness of Wage Inflation," *Brookings Papers on Economic Activity*, 1:1976, 115–59.

Labor and Output Over the Business Cycle: Some Direct Evidence

By JON A. FAY AND JAMES L. MEDOFF*

The short-run relationship between labor and output, often described in terms of short-run movements in labor productivity, is at the center of much of economic analysis. At the macroeconomic level, this relationship is a fundamental element in the aggregate supply function of an economy.[1] At the microeconomic level, it is an important element in the analysis of production and cost.

Empirical research on the short-run labor-output relationship has generally yielded estimates of the short-run elasticity of labor with respect to output that are lower than expected. Several hypotheses can explain this puzzling result. First, the technologically possible amount of labor adjustment may be smaller than suggested by the standard theory of production. Second, labor may be assigned during downturns to work other than regular production and operations tasks. Third, firms may hoard labor. And fourth, the estimated elasticities may be biased downward. The primary controversy in the literature has centered on whether firms hoard labor or whether the other explanations account for the observed cyclical movement in labor productivity.

To date, the direct evidence that is required to measure the empirical importance of the alternative hypotheses has not been available. This paper attempts to provide the necessary evidence. Section I analyzes the various hypotheses that can explain the puzzling empirical relationship between labor and output. The next section shows what information is necessary to evaluate the empirical importance of these hypotheses. In particular, it shows what data are needed to measure hoarded labor. Section III discusses the instrument that we developed to gather this information. The fourth section presents a detailed empirical analysis of the short-run relationship between output and labor in 168 plants in the U.S. manufacturing sector during recent downturns. The results of this analysis may be summarized as follows. During its most recent trough quarter, the typical plant that had a downturn paid for about 8 percent more blue-collar labor hours than were technologically necessary to meet that quarter's regular production and operations requirements. About half of this labor could be justified by the value of other work (such as additional maintenance, cleaning, training, and so on) that was completed during the trough quarter. Thus, after considering the value of any other work done by labor in excess of regular production requirements, about 4 percent of the labor paid for by the typical plant during its trough quarter should be classified as hoarded.

I. The Puzzle and Possible Solutions

Until 1960, economists generally adopted the assumptions that the technical production function exhibits decreasing average product of labor in the short run and that labor is freely variable. Taken together, these two assumptions imply that the observed short-run elasticity of labor with respect to output should be *greater than* unity, so that labor productivity should rise during down-

*Harvard University, Littauer Center, Cambridge MA 02138, and the National Bureau of Economic Research. We are grateful to Andrew Abel, Olivier Blanchard, Charles Brown, Benjamin Friedman, Robert Solow, and participants in the Labor and Macroeconomics seminars at Harvard for helpful suggestions concerning an earlier draft. David Belluck, Nicole Garris, Elizabeth Howe, Martin Van Denburgh, and especially Cliff Frazier provided valuable assistance with the research. Funding for this project was provided by a National Science Foundation grant to the National Bureau of Economic Research (no. SES 82-08-539) and by the Harvard Graduate School of Business Administration.

[1]For a good explanation of how labor hoarding affects the aggregate supply function, see Rudiger Dornbusch and Stanley Fischer (1981, ch. 11).

turns and fall during upturns. In 1960, Thor Hultgren tested these beliefs by examining the cyclical movements of output and labor for a large number of industries in the U.S. manufacturing sector. Contrary to expectation, he found "a strong tendency toward an inverse relation between hours per unit and total output" (p. 8). In 1962, using data for the U.S. economy as a whole, Arthur Okun also obtained results that were inconsistent with the "*a priori* arguments... that depressed levels of activity will stimulate productivity" (p. 103).

In 1964, Robert Solow stressed the importance to economic theory of understanding this puzzle of apparently increasing short-run returns to labor. In subsequent years, a substantial literature on the short-run relationship between labor and output developed.[2] In this literature, the phrases "short-run productivity puzzle" and "paradox of short-run increasing returns to labor" were coined to refer to the clash between a priori beliefs and empirical results. Researchers have sought to resolve this contradiction either by modifying the prior beliefs or by indicting the econometric findings.

Economists who have chosen the former alternative have questioned several key assumptions. First, some researchers have noted that the productivity puzzle could be solved by revising the traditional assumption that the technology of production is characterized by decreasing average product of labor in the short run. Several authors have suggested that the underlying technology may exhibit increasing average product of labor in the short run.[3] Many others have stopped short of accepting the hypothesis of increasing returns to labor, but have suggested that the assumed value of the technologically possible short-run elasticity of labor with respect to output should be revised downward to unity.[4]

A second solution to the short-run productivity puzzle involves examining the production process and the different types of work that are part of that process. Some of the tasks in the production process support production but do not contribute directly to output in the period during which they are performed.[5] Such tasks might include routine maintenance of equipment, overhauling machinery, cleaning, painting, training, and so on. The amount of this type of work that is performed need not be linked closely to the level of output on a period-by-period basis. Therefore, firms can postpone some of these tasks during periods when output is high.[6] For these firms, the amount of labor that is needed during a downturn is greater than the amount that would be needed if the lower level of output were permanent since additional labor is needed to perform the postponed work.[7] In this case, even if the static production function exhibits decreasing average product of labor and labor is freely variable, the percentage change in labor might not be as great as the percentage change in output during downturns.[8]

[2] For a good review of the literature through 1968, see Ray Fair (1969, ch. 2). A good review of the empirical work through 1976 can be found in Daniel Hamermesh (1976).

[3] See, for example, Okun (1962, p. 104), Frank Brechling and Peter O'Brien (1967, p. 287), and N. J. Ireland et al. (1973, p. 23). Researchers have offered a number of reasons why the technical production function might exhibit an increasing average product of labor in the short run. For example, some have argued that the short-run production function is characterized by constant proportions and that there are increasing returns to scale for various reasons. Others have argued that some portion of labor is set-up labor.

[4] See, for example, the comments by Solow on pp. 732–33 of Christopher Sims' 1974 paper. A good analysis of the attempts to reconcile the model of production with the observed labor-output relationship was presented by Robert Lucas (1970). His alternative hypothesis was later qualified by Thomas Sargent and Neil Wallace (1974).

[5] Output refers to final output or to inventories of work in process.

[6] A discussion of why firms might schedule the performance of nonproduction or support work in this way, even if labor is freely variable, can be found in our working paper version of this article (1985). (We will refer to that longer version of this article a number of times for more detailed analysis of issues that we chose not to discuss at length in this article.)

[7] Alternatively, the additional labor may perform in advance the work that will not be performed during the next peak period. Also, when we refer to a permanent level of output we mean permanent *ceteris paribus* (i.e., with the capital stock and technology fixed).

[8] We use the phrase "static production function" to describe the function that shows the minimal amount of labor needed to produce a given level of output, assuming that the firm is doing the amount of maintenance,

A third hypothesis, known as the labor hoarding hypothesis, rejects the assumption that labor is a freely variable factor. According to this hypothesis, if an estimate of the actual labor-output elasticity is less than the assumed technically possible elasticity, the difference reflects the existence of hoarded labor. In practice, an estimated elasticity of labor with respect to output that is less than one, which is the lowest value consistent with the assumption of nonincreasing returns to labor, has been treated as *prima facie* evidence of labor hoarding. This labor hoarding usually has been described in terms of "less intense effort from employees during downturns" (Martin Baily, 1978a, p. 23).[9]

Overall, the labor hoarding hypothesis has been very popular. Proponents of this hypothesis have offered a number of reasons why firms might hoard labor. These reasons were succinctly summarized very early in the literature by Okun (1962). He noted that labor hoarding might be optimal, reflecting such factors as contractual commitments that limit adjustment of labor, the transactions costs of adjustment, the value of holding an inventory of certain skills that may be needed quickly during an upturn, and the adverse effects of labor adjustment on morale. Subsequent research has elaborated considerably on several of these possible explanations.[10] On the other hand, labor hoarding could be the result of satisficing rather than optimizing behavior; or hoarding might result because managers' objective functions are not the same as those of their firms' owners.

The second major line of attack on the short-run productivity puzzle has focused not on the assumptions regarding the production process, but rather on the validity of the estimates of the actual short-run elasticity of labor with respect to output. The earliest estimates of this elasticity were significantly and substantially less than one. Several researchers felt, as Christopher Sims stated, that "some of the apparent SRIRL [short-run increasing returns to labor] reported in previous studies reflect[ed] statistical bias in estimating the labor-output relation" (p. 695), and they concentrated on deriving better estimates. Eventually, some researchers did derive estimates of the actual elasticity of labor with respect to output that were not significantly different from unity for some data sets.[11]

The researchers who have chosen the alternative of attacking the empirical evidence of short-run increasing returns to labor have given us increasingly precise estimates of the

training, and so on that would be optimal if the given level of output were permanent. The "dynamic production function," on the other hand, shows the minimal amount of labor needed to produce a given level of output assuming that postponable work is not tied to the current level of output but is instead scheduled optimally over time as output fluctuates. If firms schedule postponable work countercyclically, as described in the text, then the dynamic production function will lie above the static production function for high (relative to normal) levels of output and below it for low levels of output. For more discussion, see our working paper.

[9]For other discussions of labor hoarding in terms of intensity of labor effort see Robert Hall (1980, pp. 94–96); Dornbusch and Fischer (p. 371); Brechling (1965); S. L. Bhatia (1979); and Robert Costrell (1979). Other authors have described hoarded labor as redundant labor but have not explicitly specified decreased effort as the mechanism by which the number of hours actually worked is reduced to less than the number of hours paid for. See, for example, Douglas Greer and Stephen Rhoades (1977, p. 299); Okun (1981, p. 105); Fair (1969, ch. 3, especially pp. 37–40); Julius Chang (1983); and R. J. Ball and E. B. A. St. Cyr (1966).

[10]For example, see Costas Azariadis (1975) for a discussion of implicit contractual commitments. The potential costs of adjusting the skilled portion of the work force have been described by Walter Oi (1962) and others. Roger Miller (1971) suggested that transactions costs and skill requirements make it difficult to increase a firm's labor force rapidly as the firm comes out of a cyclical trough; therefore, a firm may want to hold a reserve of standby labor during downturns for the strategic goal of being able to expand supply rapidly when demand increases. The morale cost of labor adjustment has been mentioned often, but does not seem to have been the subject of much detailed analysis. Recently, the role of expectations of future output levels has attracted much attention. See, for example: Baily (1978b and the comments following by Benjamin Friedman, pp. 52–57); Dornbusch and Fischer (ch. 11); and Sims (pp. 698–700).

[11]For example, Sims estimated an almost proportional eventual response of man-hours of production workers to output in the aggregate durables sector, the nondurables sector, and the apparel industry. His paper offers a good analysis of the econometric issues that are involved. And John Tatom (1980) claimed to have found evidence of diminishing returns to labor in the aggregate U.S. private business sector.

actual elasticity of labor with respect to output. However, their efforts also point out the fundamental shortcoming of the empirical research to date on short-run productivity movements: data on the observed responsiveness of labor to output do not contain the detailed information that is necessary to directly evaluate the alternative explanations of that responsiveness. Estimates of the labor-output elasticity that were less than unity may have prompted the formulation of the hypotheses regarding technology, other work, and hoarding, but new estimates that are equal to or greater than unity do not constitute sufficient evidence to dismiss these hypotheses. We still do not know whether firms hold labor in excess of regular production requirements, nor do we know whether firms hoard labor. In short, there are a number of hypotheses that can explain the observed short-run relationship between labor and output, but in order to measure the empirical importance of these hypotheses, we must gather additional data.

II. The Necessary Data

Consider a firm that has experienced a cyclical decline in real shipments from S_N (the normal level, based on long-term trend and seasonal factors) to the trough level S_T (the level that is furthest below normal). Suppose that in response to this decline in shipments, the firm decreases its output from Q_N (the normal level) to Q_T (the level during the same period as the shipments trough). When output is cut to Q_T, the firm reduces the amount of labor it is paying for from L_N (the normal level of labor input) to L_T (the amount of labor paid for during the trough period).[12] Generally speaking, this is the information that is publicly available. However, to measure the importance of the hypotheses that were presented in the previous section, we need more data. This section describes in detail the necessary data.

We begin by noting that L_T can be separated into two parts, L_R and L_O. The term L_R represents the labor that is assigned to the direct production tasks that are required to produce Q_T and to the indirect tasks that would be required to support production if Q_T were the permanent level of output. L_O represents the labor that is assigned to all tasks that would not optimally be scheduled if Q_T were permanent. Thus, L_O might include postponable work, as described in Section I, and it might also include cyclical make-work. Hereafter, we will refer to L_R as the labor that is assigned to regular production work and L_O as the labor that is assigned to other work.[13] By definition, $L_T = L_R + L_O$.

Next, we can analytically separate both L_R and L_O into a worthwhile component and a hoarded component. Consider first the labor that is assigned to regular production work, L_R. Let L_C represent the minimal amount of labor that is technologically needed to perform the regular production work, assuming that labor effort stayed at its normal level if technologically possible.[14] If $L_R > L_C$, then more labor is assigned to regular production work than is technologically necessary. If we let $L_{HR} = L_R - L_C$, then L_{HR} represents the hoarded component of the labor that is assigned to regular production work.

The component L_{HR} might manifest itself as decreased effort on regular production assignments.[15] Alternatively, suppose that a firm wants to protect its most skilled or most

[12] We will focus on the adjustment of labor given the level of output. We will not deal explicitly with the question of why the firm adjusts its output and inventories in response to a cyclical decline in shipments.

[13] The abbreviated descriptions of L_R and L_O will be used for convenience. Bear in mind, however, that L_R includes the labor doing the indirect support tasks that would be optimally scheduled if Q_T were permanent.

[14] Thus, L_C represents the minimal amount of labor that is technologically necessary to produce Q_T and to perform the support work that would be optimal if Q_T were permanent.

[15] Decreased effort could take many forms. Employees could work the number of hours they are paid for, but work less intensely than normal. Or, they could work with normal intensity while working, but work fewer hours than they are paid for. Also, the decreased effort might be enjoyed by all employees or by just a few. Not all forms will be readily apparent to an outside observer.

senior employees by allocating the required production work to these employees and concentrating the layoffs and hours cuts among other workers. The workers who are retained may be less efficient in the short term, even if their effort remains normal, because they are performing jobs that they are not used to doing. In this case, the firm pays for more regular production labor than is technologically necessary ($L_{HR} > 0$), even if effort remains normal.[16]

The labor L_O can similarly be separated analytically into a worthwhile component and a hoarded component. Let L_W represent the greatest amount of labor that could be assigned to other work, subject to the constraint that the marginal revenue product of this labor must be greater than or equal to the wage rate, and assuming that effort stays at its normal level. Roughly speaking, L_W is the minimal amount of labor needed to perform the worthwhile other work. If we let $L_{HO} = L_O - L_W$, then L_{HO} represents the hoarded component of the labor that is performing other work.

The component L_{HO} can take many forms. More labor may be assigned to other work than is technically necessary given the amount of other work that is being accomplished. Alternatively, some of the other work may be make-work that is inherently of little value, regardless of how hard or efficiently employees work at it.

With the information described above, we can measure the amount of labor that is paid for in excess of regular production requirements. $L_E = L_T - L_C$ measures the amount of labor that the firm is paying for in excess of the amount technologically necessary to perform the regular production and operations tasks needed to produce Q_T.

We can also measure the amount of hoarded labor. To do this, we must subtract the amount of labor accounted for by worthwhile other work from the amount of labor in excess of regular production requirements. Thus, $L_H = L_T - L_C - L_W$, where L_H stands for the total amount of hoarded labor. Alternatively, we can measure labor hoarding in a manner that emphasizes that hoarding can occur both on regular production assignments and also on other assignments. Specifically, we can measure hoarded labor as $(L_R - L_C) + (L_O - L_W) = L_{HR} + L_{HO} = L_H$. Measuring L_{HR} and L_{HO} separately gives us information on how firms hoard labor, as well as on how much they hoard. A more detailed description of the ways in which labor is hoarded would require additional information, such as data on effort.

III. Gathering the Data

To gather the necessary data, we surveyed plant managers regarding their most recent completed cyclical downturns.[17,18] The first section of the questionnaire began with several definitions, including the following:

BUSINESS DOWNTURN: A period when the *real* shipments of your products have declined to below normal levels. That is, a cyclical business downturn is when real shipments have fallen below what you could normally expect on the basis of regular seasonal factors and long-term (10–20 year) trends.

TROUGH: The quarter of the downturn when real shipments were furthest below normal.

The first questions in the section asked when the various points in the cycle (peak, normal, trough, and return to normal) occurred. The next questions asked by what percentage real shipments and production were below normal during the trough quarter; these questions provide data on $(S_N - S_T)/S_N$ and

[16] Note also that this hoarded labor would not constitute a reserve of labor that could be used to increase output without increasing payroll costs during a subsequent upturn.

[17] A copy of the questionnaire that is described in this section is available from the authors.

[18] Each respondent was asked to describe the most recent completed downturn unless an ongoing downturn was the only one that he or she had experienced at his or her current plant. Thirty-two percent of the respondents answered for an ongoing downturn; they were instructed to consider the quarter in which they answered the questionnaire to be the trough quarter.

$(Q_N - Q_T)/Q_N$. (All of the survey data were gathered in percentage form.)

The second section gathered information about labor adjustment.[19] To measure the actual amount of labor adjustment we asked:

By what percentage was the total number of blue-collar hours paid for *actually* below normal during the trough quarter (when real shipments were furthest below normal)?

__percent below normal at the trough.

We interpret the answer to this question as a measure of $(L_N - L_T)/L_N$.

The inquiry regarding the technologically possible adjustment began with the following definition:

MAXIMUM REDUCTION TECHNICALLY POSSIBLE: The next question asks how much the total number of blue-collar hours paid for could have been reduced if the *only* consideration had been the technical requirements of your plant's production process. In other words, it asks what the maximum possible reduction in total blue-collar hours could have been if your plant had produced the same level of output as it did during the downturn and:

—such concerns as employee morale, union contract restrictions, company image, the possibility that skilled workers would not return from layoff, the protection of senior workers, uncertainty about future demand, and so on *were not considered*;

—necessary (but not additional) maintenance and training *were* continued;

—blue-collar employees had worked at the same level of effort as in *normal* times.

[19] We will use the questions about labor adjustment to measure the variables described in Section II. As with all empirical research, our interpretations of the answers to these questions require a number of assumptions. Our purpose in designing a survey specifically to gather information on labor adjustment was to minimize the number and severity of the assumptions that had to be made to measure the relevant variables. We believe that the assumptions that must still be made in interpreting the survey data are all very reasonable. We have analyzed these assumptions at length in our working paper (see especially the footnotes to Sections III and IV).

The question that followed this definition sought information on the technically possible amount of labor adjustment, or $(L_N - L_C)/L_N$:

If blue-collar employment and hours had been cut back to the minimum levels *technically* necessary to maintain your plant's actual level of output (and necessary operations) during the downturn, by what percentage would the total number of blue-collar hours paid for have been below normal during the trough quarter?

__percent below normal at the trough if we had made the maximum reduction technically possible.

To gather information on the value of any other work that was assigned to employees in addition to regular production and operations tasks, we began by asking what types of other work had been assigned:

During the trough quarter of the downturn, what types of work were assigned to your blue-collar workers in addition to or instead of their normal assignments? What types of work were intentionally deferred during normal times and assigned during the downturn? Please check all that apply.

__None (Please skip to the next question)

__Maintenance of equipment

__Overhaul of equipment

__Painting of plant and equipment

__Cleaning

__Reworking output which failed inspection

__Receiving and/or giving training

__Other (please specify).

Next, we asked the manager to quantify how much blue-collar labor was assigned to this other work. The question that we used to measure L_O/L_T was:

During the trough quarter, approximately what percentage of your total blue-collar hours was spent on the additional and deferred assignments mentioned in the previous question?

__percent on additional and deferred assignments.

We also need to know how much of this other work was worthwhile and how much was not. Therefore, we asked the following question to gather data on L_{HO}/L_O:

Roughly what percentage of the hours spent on additional and deferred assignments during the trough quarter were worth less to your company than hours spent by the same employees on their regular assignments during a normal quarter? In determining the "worth" of an assignment, compare the costs of the assignment (including wages as well as materials, wear and tear on equipment, etc.) to the present and future benefits (revenue from reworked items, longer equipment lifetimes, fewer breakdowns, better trained workers, etc.).

__percent.

In order to gather more information on how labor is hoarded, we also asked about effort. Specifically, we asked each respondent to rate average blue-collar worker effort as much lower than normal, slightly lower than normal, the same as normal, slightly greater than normal, or much greater than normal for each of three points in the business cycle (when shipments were just starting to fall from their peak level, when they were just beginning to fall below normal, and when shipments were at their trough level).

The comments of many managers and executives guided both the direction and the wording of the questions. After the final questionnaire was completed, it was pretested for validity. During pretests, we met with managers who had answered the questionnaire in order to verify the managers' interpretations of the questions and to verify that the managers' answers accurately reflected their plants' operating data.

After the pretests, we surveyed a sample of U.S. manufacturing establishments. In August of 1981, we sent the questionnaire to 498 companies that were chosen randomly from the *Newsfront* list of the largest 2684 manufacturing enterprises. Our mailing was directed to vice presidents in charge of manufacturing, operations, or human resources management, or to the company presidents. These executives were asked to forward the two questionnaires they received to two of their plant managers (if they had at least two plants), unless the questionnaire recipients felt that they should be the respondents. In April 1982, we sent a second wave of questionnaires directly to 1000 plant managers whose establishments had been selected randomly from the Marketing Economics Institute listing of manufacturing plants with 100 or more employees. In October of 1982, a followup questionnaire was sent to each plant manager in the second wave from whom we had received no response.

A total of 326 questionnaires were returned to us. Of these, 41 were returned because they were undeliverable, 10 were returned by managers whose plants did not fit the sample criteria, and 14 were returned by managers who declined to participate in the survey.[20] Of the 261 questionnaires that were returned with answers, 15 did not have answers to one or more questions and another 4 were not usable for other reasons.[21] Of the 242 properly completed questionnaires, 68 were from managers who had never witnessed a downturn at their current plants and 6 were from managers whose plants had

[20] The questionnaire included the following instructions:

Please answer this survey only if: (1) you are employed by a private sector, for-profit firm whose principal activity is manufacturing; and (2) you are a manager at the plant or division level who is directly involved in managing production, operations, and human resources.

[21] One questionnaire was dropped from the data set because the respondent gave answers based on his predicted response to an anticipated downturn. Another was excluded because the respondent indicated that his answers pertained to seasonal fluctuations. A third observation was dropped because the respondent apparently made computational errors. A fourth observation was excluded because including it caused the estimates of the elasticities in Table 4 to be misleading since the values for this observation were so far from the mean for some variables; see fn. 37 for more discussion. Excluding this observation did not qualitatively affect the estimates of excess labor or hoarded labor presented in Tables 2, 3, or 5. An attempt was made to obtain missing information on incomplete questionnaires and to investigate all seemingly inconsistent or invalid responses.

TABLE 1—COMPARISON OF SURVEY SAMPLE TO U.S. MANUFACTURING SECTOR[a]

Characteristic	COM	Survey ($N=239$)	Survey ($N=168$)
Percentage of Establishments with the Following Characteristics:			
Durables Manufacturer	58	59	58
More than 20 Employees	32	99	100
Northeast (*COM* region definition)	27	24	25
North Central	26	36	40
South	27	29	24
West	20	10	10
Percentage of Production Employees in Establishments with the Following Characteristics:			
Durables Manufacturer	59	61	63
More than 20 Employees	93	100	100
Northeast	24	17	16
North Central	32	48	53
South	31	28	23
West	13	7	7
Percentage of Production Employees Covered by Collective Bargaining Aggreements	61	65	68

[a]Percentages may not add properly due to rounding error.

not cut production. Thus, 168 questionnaires were available for analysis.

We examined the representativeness of the group of respondents in several ways. First, we compared the characteristics of the survey respondents' plants to the characteristics of U.S. manufacturing establishments as reported in the *1977 Census of Manufactures* (*COM*). This comparison is summarized in Table 1.[22] The first column of figures shows the characteristics of establishments in the *COM*, the second column describes the plants of the managers who returned a properly completed questionnaire, and the third column gives the characteristics of the plants that will be analyzed in the next section.[23] Along most dimensions, the survey sample is representative of the U.S. manufacturing sector. A notable exception is establishment size: the survey sample consists almost entirely of plants with more than 20 employees. However, as the table shows, larger establishments account for a very large percentage of the employment of production workers in the U.S. manufacturing sector. For example, *Newsfront* estimated that the companies on its list accounted for more than 90 percent of total manufacturing employment.

Second, we compared the estimates of labor in excess of regular production requirements and of hoarded labor from the first survey of plants on the Marketing Economics Institute list to the estimates from the followup survey to see if the data for initial nonrespondents differed significantly from those of initial respondents. We also compared the estimates for plants from the *Newsfront* sample to the estimates for all plants from the Marketing Economics Institute sample to see if the survey technique was important. These tests revealed no significant differences.

Finally, we compared the survey data on actual labor and output changes to the comparable data from previous studies. As reported in the next section, the survey esti-

[22]The data on percentage covered by collective bargaining for all manufacturing are from Richard Freeman and Medoff (1979).

[23]The statistics in the second column are based on the data for 239 respondents because 3 of the 242 questionnaires that contained all of the data needed for the analysis of labor adjustment did not include some item of background data on plant characteristics.

mates of the actual (observed) labor-output and labor-shipments elasticities are very similar to the estimates derived from publicly available data and reported in earlier research.

IV. The Direct Evidence

The summary of the survey data begins in Table 2. The first column of summary statistics presents the mean and the standard error (*SE*) of the mean of each variable for all 168 observations. The second data column presents the mean and the *SE* of the mean of each variable for the 78 respondents who did not assign any work other than regular production tasks during the trough quarter of the downturn.[24] The third column of results presents the summary statistics for the 90 respondents who did assign other work during the trough quarter. The following discussion focuses on the results for all respondents.

Rows 1–3 of Table 2 describe the characteristics of the respondents' downturns. It took four quarters on average for shipments to fall from normal to the level furthest below normal, and shipments were 30 percent below normal on average during the trough quarter.[25] The mean decline in production was 31 percent for the whole sample; thus, on average, the establishments in the sample appear to have been drawing down inventories very slightly during the trough quarter.

The analysis of the patterns of labor adjustment begins in row 4. On average, it would have been technologically possible to have reduced the total number of blue-collar hours paid for during the trough quarter to 28 percent below normal, given the level of production. The actual reduction in total blue-collar hours was 23 percent, on average. The difference between the technically possible labor reduction and the actual reduction measures the amount of labor that was paid for in excess of regular production requirements. As shown in row 6, the mean amount of labor in excess of the minimum needed for regular production tasks equalled about 6 percent of the amount of labor normally paid for. Thus, starting from normal, the typical plant could have cut back 6 percent more than it did and still have completed the regular production work of the trough quarter.[26] The mean amount of labor in excess of regular production needs was significantly greater than zero at the .05 level for the entire sample and for both subsamples.[27] Thus, the survey data indicate that the typical manufacturing establishment does pay for a nontrivial amount of blue-collar labor in excess of regular production requirements during the trough quarter of a downturn.[28]

[24] The subsample of respondents who did not assign other work includes those who replied "None" when asked what types of other work they had assigned, and those who replied that they assigned zero percent of their labor to other tasks.

[25] The length of a downturn was calculated as the quarter of the trough minus the quarter in which shipments fell below normal plus one. See fn. 18 for the definition of the trough quarter. If the respondents who had never experienced a business downturn are included in the sample and assigned a shipments reduction of zero, the mean percentage below normal was 21.

[26] This does not necessarily mean that the typical plant, once it had reached the trough quarter, could have further reduced its labor input by an amount equal to 6 percent of its normal labor input. Even if it is not performing other work, labor in excess of regular production requirements does not necessarily constitute a reserve of redundant labor. (Also, note that some figures may not add properly due to rounding error.)

[27] A one-tailed test is appropriate in light of the post-1960 literature, since the debate in the literature has centered on whether labor hoarding occurs. We will report in the text the results of significance tests that were performed with means and standard errors that were calculated to several decimal places; readers who perform tests using the rounded figures from the tables will sometimes obtain slightly different results due to rounding error.

[28] The distribution of this variable for the subsample of plants that did not assign other work would appear to be very close to a normal distribution centered at about 5 percent, except for a large spike at zero. Because of the position of the spike to the left of the mean of an otherwise normal distribution, we believe that many respondents answered that the technically possible adjustment was equal to the actual adjustment even though there was actually a small difference between the two. Thus, in our opinion, the sample mean of 3 percent probably underestimates the true population mean. A spike does not appear in the distribution for the subsample of plants that did assign other work. Accordingly, there is only a small spike at zero in the distribution for the complete sample.

TABLE 2—SURVEY RESULTS[a]

Variable[c]		Mean of Variable for[b] All Respondents ($N=168$)	Respondents Who Did Not Assign Other Work ($N=78$)	Respondents Who Did Assign Other Work ($N=90$)
1) Quarters from Normal to Trough		4 (0)	4 (0)	4 (0)
2) Percentage Reduction in Shipments	$(S_N - S_T)/S_N$	30 (1)	32 (2)	29 (2)
3) Percentage Reduction in Production	$(Q_N - Q_T)/Q_N$	31 (2)	33 (2)	30 (2)
4) Technically Possible Percentage Reduction in Total Blue-Collar Hours	$(L_N - L_C)/L_N$	28 (1)	28 (2)	29 (2)
5) Actual Percentage Reduction in Total Blue-Collar Hours	$(L_N - L_T)/L_N$	23 (1)	25 (2)	21 (2)
6) Percentage of Normal Labor Hours that Technically Could Have Been Eliminated But Were Not	$(L_T - L_C)/L_N$	6 (1)	3 (2)	8 (2)
7) Types of Other Work Assigned During Trough Quarter (Yes = 100)				
No Other Work		46	100	0
Maintenance of Equipment		36	0	67
Overhaul of Equipment		24	0	46
Painting		35	0	64
Cleaning		37	0	69
Reworking Output		17	0	32
Training		18	0	34
Other		11	0	21
8) Percentage of Normal Hours Assigned to Other Work	L_O/L_N	5 (1)	0 (0)	8 (1)
9) Percentage of Normal Hours Assigned to Worthwhile Other Work	L_W/L_N	3 (0)	0 (0)	5 (0)
10) Hoarded Labor as a Percentage of Normal Hours	L_H/L_N	3 (1)	3 (2)	3 (2)
11) Hoarded Labor on Regular Production Work, as a Percentage of Normal Hours	L_{HR}/L_N	1 (1)	3 (2)	−1 (2)
12) Hoarded Labor on Other Work, as a Percentage of Normal Hours	L_{HO}/L_N	2 (1)	0 (0)	3 (1)

[a]See Table 1.
[b]The standard errors are shown in parentheses.
[c]The algebraic definitions in the second column must be multiplied by 100 to obtain the percentage forms of the variables.

For the establishments that did not assign any other work, the labor in excess of regular production needs was all hoarded labor. But for the other plants we must consider the value of the other work that was completed during the trough quarter. Row 7 shows the percentage of respondents who assigned each type of other work.[29] Row 8 shows that 5 percent of the typical plant's normal blue-collar labor had been assigned to other work by the trough quarter. Row 9 summarizes the data concerning what portion of the labor that was assigned to other work should be considered worthwhile; specifically, the variable described in that row measures L_W as a percentage of the normal labor input. As shown in row 9, the mean of this variable was 3 percent. Thus, of the 5 percent of the labor that was assigned to other work, on average, 3 percentage points could be justified by the value of the other work.

The amount of labor that can be accounted for by the value of other work must be subtracted from the amount of labor that was held in excess of regular production requirements in order to arrive at a measure of hoarded labor. This calculation was performed for each plant, and the results are summarized in row 10, which shows the amount of blue-collar labor that was hoarded. As shown in that row, the average plant hoarded an amount of labor equal to 3 percent of its normal labor input. The mean amount of hoarded labor was significantly greater than zero at the .05 level for the complete sample. It was also significantly greater than zero at the .05 level for the first subsample. For the second subsample, the mean amount of hoarded labor barely missed being significantly greater than zero at the .05 level (it was significantly greater than zero at the .0509 level). In over half of all plants, the amount of hoarded labor was between 0 and 8 percent of the amount of labor normally paid for.

The last two rows of Table 2 summarize the survey evidence regarding the forms in which labor hoarding occurs. As shown in row 11, the amount of hoarded labor on regular production assignments was equal to 1 percent of the amount of labor normally paid for, on average. The amount of hoarded labor doing other work equalled 2 percent of the normal amount of labor input, on average, as shown in the final row of the table.

Additional understanding of the ways in which labor is hoarded can be gained from the survey data regarding effort. Because the measure of effort is crude, the evidence regarding effort must be viewed as only suggestive. Bearing this in mind, the evidence does not suggest that hoarding predominantly takes the form of decreased effort.[30] Thus, mechanisms other than decreased effort appear to be important in turning paid-for hours into hoarded hours.

Thus far, the discussion of Table 2 has focused on the results for the complete sample. It is also interesting to note the differences between the plants that assigned other work and those that did not, even though these differences are generally not statistically significant due to the smaller sizes of the subsamples. The plants that did assign other work held more labor in excess of regular production requirements. However, once the value of the other work is accounted for, the amount of hoarding in the two sub-

[29] The "Other" types of other work that were assigned by 11 percent of all plants included: "new product development," "new methods," "manufacturing merchandising aids," and other "discretionary" and "less profitable work [that] needed to be done."

[30] The evidence regarding the relationship between hoarding and change in effort is summarized in the following tabulation. The numbers indicate the number of questionnaires that fell in each cell. The three categories of change in effort refer to the change in effort from normal to trough. The four columns correspond to plants in which there was: hoarded labor on regular production assignments only; hoarded labor on other assignments only; hoarded labor on regular and on other assignments; and no hoarded labor. By comparing the pattern of effort change across different columns, we can see very roughly whether each type of hoarding is generally accompanied by a decrease in effort. The percentages sum to 100 (plus or minus rounding error) within columns.

	$L_{HR} > 0$	$L_{HO} > 0$	$L_{HR} > 0$ $L_{HO} > 0$	No Hoarding
Increase	10(22%)	12(41%)	12(30%)	23(43%)
No Change	20(44%)	13(45%)	12(30%)	17(31%)
Decrease	15(33%)	4(14%)	16(40%)	14(26%)

TABLE 3—DECOMPOSITION OF LABOR HOURS PAID FOR DURING THE TROUGH QUARTER[a]

Variable	All Respondents ($N=165$)	Respondents Who Did Not Assign Other Work ($N=77$)	Respondents Who Did Assign Other Work ($N=88$)
	Mean of Variable for[b]		
Hours Paid for at Trough (L_T)	100 (0)	100 (0)	100 (0)
of Which:			
Hours Spent on Regular Production Work (L_R),	94 (1)	100 (0)	89 (1)
of Which:			
Hours Technically Necessary for Regular Production Work (L_C)	92 (1)	96 (2)	89 (2)
Hoarded Hours on Regular Production Work (L_{HR})	2 (1)	4 (2)	0 (2)
Hours Spent on Other Work (L_O),	6 (1)	0 (0)	11 (1)
of Which:			
Worthwhile Hours on Other Work (L_W)	3 (0)	0 (0)	6 (1)
Hoarded Hours on Other Work (L_{HO})	2 (1)	0 (0)	4 (1)
Hours in Excess of Regular Production Requirements ($L_T - L_C$)	8 (1)	4 (2)	11 (2)
Total Hoarded Hours ($L_H = L_{HR} + L_{HO}$)	4 (1)	4 (2)	5 (2)

[a]All variables are measured as a percentage of the total number of blue-collar hours paid for during the trough quarter (L_T). Percentages may not add properly due to rounding error.

[b]See Table 2.

samples is the same. Among the establishments that did not assign other work, all of the labor hoarding occurred on regular production assignments, by definition; among the establishments that did assign other work, the pattern is nearly the opposite.[31]

Overall, the survey data indicate that the typical U.S. manufacturing plant does hoard blue-collar labor during the trough quarter of a downturn. The data also indicate that hoarding is a complex phenomenon—hoarded labor takes a number of the different forms that were discussed in Section II.

Table 2 began with the amount of blue-collar labor that was normally paid for and showed what had happened to that labor by the trough quarter of the downturn; it showed how much labor was eliminated, how much was reassigned to other work, and so on. In Table 3, we begin with the total amount of blue-collar labor that was paid for during the trough quarter as the base, and then provide a complete and explicit decomposition of the labor paid for during the trough quarter.[32]

[31]The decomposition of hoarded labor into L_{HR} and L_{HO} is subject to a number of errors that were discussed in our working paper. Additional information might lead to a slight modification of the decomposition. However, none of these errors affects the estimates of the total amount of hoarded labor, L_H.

[32]Since the variables in Table 3 are all denominated by L_T, the 3 plants that reduced their blue-collar labor to zero could not be included in this analysis. Therefore, the samples in Table 3 differ from the samples in Tables 2, 4, and 5.

TABLE 4—ARC ELASTICITIES OF LABOR WITH RESPECT TO OUTPUT[a]

Variable	Mean of variable for[b]: All Respondents ($N=168$)	Respondents Who Did Not Assign Other Work ($N=78$)	Respondents Who Did Assign Other Work ($N=90$)
1) Actual Elasticity of Total Blue-Collar Hours with Respect to Output	.82 (.05)	.86 (.07)	.78 (.07)
2) Technically Possible Elasticity of Regular Production Labor with Respect to Output	1.17 (.08)	1.10 (.12)	1.22 (.09)
3) Difference Between the Technically Possible and Actual Elasticities	.35 (.07)	.23 (.09)	.44 (.10)
4) Actual Elasticity, Corrected for Adjustment of Labor into Worthwhile Other Work	.96 (.06)	.86 (.07)	1.05 (.08)
5) Difference Between the Technically Possible and Corrected Actual Elasticities	.20 (.06)	.23 (.09)	.17 (.09)

[a]See Table 1.
[b]See Table 2.

In the full sample, the typical establishment assigned 94 percent of the labor that it paid for during the trough quater to regular production work. Of those 94 percentage points, 92 were technologically necessary given the level of production, and the remaining 2 percentage points were hoarded. Six percent of the labor that was being paid for by the typical plant was assigned to other work. Of those 6 percentage points, about 3 could be justified by the value of the other work, and about 2 percentage points were hoarded.[33] The decomposition of paid-for labor for each of the two subsamples is presented in the two right-hand columns of the table.

The last two rows of Table 3 summarize how much of the labor paid for during the trough quarter was in excess of regular production requirements and how much was hoarded. On average, 8 percent of the blue-collar hours that were paid for during the trough quarter were not technologically necessary to perform the regular production work of that quarter. Four percent of the hours that were paid for were hoarded. The mean amount of labor in excess of regular production requirements and the mean amount of hoarded labor were both significantly greater than zero at the .05 level for the full sample and for both subsamples.

Since much of the short-run productivity literature has discussed labor adjustment in terms of elasticities, it is useful to analyze the survey data in these terms. Table 4 summarizes various arc elasticities of total blue-collar hours with respect to output, while Table 5 presents comparable elasticities with respect to shipments.

As shown in Table 4, the actual elasticity of total blue-collar hours with respect to output was .82 in the typical plant.[34] The technologically possible elasticity of total blue-collar hours with respect to output was

[33]The figures do not add properly because of rounding error.

[34]The actual arc elasticity was calculated as

$$[(L_N - L_T)/(L_N + L_T)]/[(Q_N - Q_T)/(Q_N + Q_T)]$$

1.17 on average.[35] Our primary concern in this article is not with these elasticities but with the difference between them. The difference between the technically possible and actual elasticities reflects the extent to which a plant held labor in excess of regular production requirements. As shown in row 3, the difference between the technically possible and the actual elasticities was .35 for the typical plant. The mean difference was significantly greater than zero at any customary level of confidence for the complete sample and for both subsamples.

In order to derive a measure of hoarding, we must take into account the value of the other work that was performed during the trough quarter. The technically possible elasticity reflects the largest possible reduction that can be made in the labor available for regular production work. However, the actual elasticity in row 1 reflects only that portion of the decrease in labor available for regular production work that is due to complete elimination of labor from the plant (via layoffs, hours cuts, or attrition). The actual elasticity does not capture the decrease in labor available for regular production work that is due to the "adjustment" of labor into worthwhile other work. In order to account for the value of other work, we constructed a corrected actual elasticity. The percentage change in labor that is used in calculating this corrected elasticity is made up of two components: the percentage of total blue-collar hours that were eliminated via layoffs, hours cuts, and attrition; and the percentage of normal hours that can be justified by the value of worthwhile other work.[36]

As shown in row 4 of Table 4, the mean value of the corrected actual elasticity of total blue-collar hours with respect to output was .96 for the full sample. As shown in row 5, the difference between the technically possible elasticity and the corrected actual elasticity was .20, on average. This difference reflects the existence of labor that could not be accounted for by regular production requirements or by the value of other work; more simply, the difference reflects the existence of hoarded labor. The mean difference was significantly greater than zero at the .05 level for the full sample and for both subsamples; moreover, judging by the standards in the literature, the mean difference was substantial.[37]

Table 5 examines analogous elasticities with respect to shipments. In general, the summary statistics in Table 5 are very similar to those in Table 4. The mean elasticities with respect to shipments (rows 1; 2; 4) were slightly larger than the corresponding elasticities with respect to output, as expected, since the mean percentage change in shipments

[35] The technologically possible elasticity was calculated as

$$[(L_N - L_C)/(L_N + L_C)] \div [(Q_N - Q_T)/(Q_N + Q_T)]$$

The survey data provide support, albeit not strong support statistically speaking, for the hypothesis that the technology of the U.S. manufacturing sector exhibits diminishing average product of labor, on balance, in the short to intermediate run. However, the mean value of 1.17 should be kept in perspective. Twenty percent of the survey respondents stated that the technically possible elasticity in their plants was 1.47 or greater, and another 20 percent stated that it was .55 or less. Any statement about the technology of the manufacturing sector as a whole involves important aggregation issues. Also, it should be noted that our estimates probably correspond to intermediate-run elasticities, since capital probably varied to some extent during the four quarters from normal to trough.

[36] The corrected actual elasticity was calculated as

$$[(L_N - (L_T - L_W))/(L_N + (L_T - L_W))] \div [(Q_N - Q_T)/(Q_N + Q_T)].$$

An alternative approach would have been to correct the actual elasticity by adjusting the percentage change in output to properly capture the value of the other work.

[37] As mentioned earlier, one extreme outlier was dropped from the analysis because its inclusion caused some of the figures in Table 4 to be misleading. The respondent in question made a moderate labor adjustment but only a small cut in output in response to a moderate drop in shipments. Since the actual elasticity was denominated by a figure close to zero, the elasticity was very large. The technically possible elasticity, however, was much closer to average. As a result, the differences between the technically possible and the actual and corrected actual elasticities were sizeable negative numbers. Including this observation would thus have lowered the mean differences in rows 3 and 5. The figures in Table 5 would not have been altered substantially by its inclusion, since denominating by the moderate drop in shipments did not cause a similar problem.

TABLE 5—ARC ELASTICITIES OF LABOR WITH RESPECT TO SHIPMENTS[a]

Variable	Mean of variable for[b] All Respondents ($N=168$)	Respondents Who Did Not Assign Other Work ($N=78$)	Respondents Who Did Assign Other Work ($N=90$)
1) Actual Elasticity of Total Blue-Collar Hours with Respect to Shipments	.88 (.06)	.90 (.08)	.86 (.10)
2) Technically Possible Elasticity of Regular Production Labor with Respect to Shipments	1.22 (.08)	1.11 (.12)	1.31 (.12)
3) Difference Between the Technically Possible and Actual Elasticities	.34 (.07)	.21 (.10)	.45 (.11)
4) Actual Elasticity, Corrected for Adjustment of Labor into Worthwhile Other Work	1.03 (.07)	.90 (.08)	1.14 (.11)
5) Difference Between the Technically Possible and Corrected Actual Elasticities	.19 (.07)	.21 (.10)	.17 (.10)

[a]See Table 1.
[b]See Table 2.

was slightly smaller than the mean percentage change in output. The differences that reflect the existence of labor in excess of regular production requirements and hoarded labor (rows 3 and 5, respectively) were very similar to the differences in Table 4, as expected.

Since much of the previous empirical research has used shipments data, the figures in Table 5 are probably the estimates that are most comparable to earlier quantitative work. It is therefore worth comparing the picture painted by the survey results to the pictures painted by earlier research. Previous studies have had available only the type of information summarized in row 1 of the table. Looking at row 1, we can see that the mean value of the actual elasticity of blue-collar labor with respect to shipments for the survey sample is very close to most recent estimates of this elasticity using publicly available data. Moreover, the mean of the uncorrected actual elasticity was not significantly different from unity at the .05 level in the full sample or in either of the subsamples.[38] Nonetheless, we have seen that the mean amount of labor in excess of regular production requirements and the mean amount of hoarded labor were significantly greater than zero in the full sample and in both subsamples. Therefore, testing whether an estimated actual elasticity is significantly different from unity is not equivalent to testing whether firms hold labor in excess of regular production requirements or whether firms hoard labor—we need the additional evidence that is provided by the survey.[39]

[38]This is the test that has been used most often in the literature.

[39]Fair (1985) provides a detailed comparison of the survey estimates of hoarding and estimates of hoarding derived with publicly available data using his macro model of the economy. Because the collection of original survey data is expensive, we agree that it is important to determine how publicly available data can be used to accurately measure the complex process of labor adjustment. Fair's comparison is based on a preliminary draft

V. Conclusion

This article has reported the results from a survey that collected direct evidence regarding labor adjustment in U.S. manufacturing establishments during business downturns. The survey data indicate that the typical plant paid for 8 percent more blue-collar labor hours than were technologically necessary to perform regular production work during the trough quarter of its most recent downturn. Not all of the labor in excess of regular production requirements should be classified as hoarded, since some of it was performing worthwhile other work. After taking into account how much of the labor could be justified by the value of the other work, 4 percent of the blue-collar hours paid for by the typical plant during its trough quarter should be classified as hoarded.

Direct evidence that firms pay for a large amount of labor in excess of regular production requirements and hoard a substantial amount of labor during downturns is important for several reasons.[40] First, the evidence indicates that a sizeable portion of the swings in productivity over the business cycle is, in fact, the result of firms' decisions to hold labor in excess of regular production requirements and to hoard labor. If the amount of labor hoarding varies with the size, duration, or other characteristics of the business cycle, then the aggregate supply response of the economy will probably be different for different types of cycles.[41] Moreover, the form, as well as the total amount, of hoarding is important, since not all forms of hoarding provide a reserve of redundant labor for the subsequent recovery. Also, the impact of a given change in output on the amount of labor that is paid for may depend on the amount of worthwhile other work that is available. The amount of worthwhile work that is available may in turn depend on the length and magnitude of the downturn and on the proximity and characteristics of previous downturns. In short, different changes in product demand may result in different levels of productivity change, and a given level of productivity change may have a number of different interpretations. Correctly interpreting a productivity change is important in understanding the behavior of the economy during the subsequent recovery.[42]

Second, the survey evidence indicates that labor is not always fully utilized in producing measured ouput. During downturns, the utilization rate of the labor that is assigned to regular production assignments is lower than normal in many plants due to hoarding on those assignments. Moreover, some labor is assigned to work that is probably not captured by measured output. Research on the production function should take this evidence into account.[43] Third, the survey evi-

of this article; some of the results he cites differ slightly from the final results presented here.

[40] It is worth emphasizing that our measure of hoarding is made relative to normal. We assume that the cyclical hoarding that we measure is zero on average during normal times. Also, the survey evidence pertains only to blue-collar workers in the U.S. manufacturing sector. However, as Sims (p. 725) and others have pointed out, hoarding is probably at least as great among most other kinds of workers and in most other sectors. Therefore, we believe that it is reasonable at this stage to consider the implications of the survey results for the whole economy.

[41] The importance of considering the characteristics of the downturn before invoking Okun's Law was noted by Okun himself in his original article (1962, p. 100).

[42] For example, if a large amount of labor is assigned to worthwhile other work that is not properly captured in the publicly available output data, then there will be a large observed drop in productivity. The interpretation of this decline should be different from the interpretation of a decline that is due to labor hoarding. Also, if the productivity change is due to hoarding, the subsequent pattern of change will depend on the form of hoarding.

[43] A number of theoretical models (for example, the one developed by Lucas) allow the rate of capital utilization to vary but assume that the firm "is scheduling labor optimally (or...is 'on' its short-run cost curve...)" (Lucas, p. 25). If both capital and labor are frequently underutilized, identifying the parameters of the technical production function becomes problematic. The issue of postponable work is similarly important. Several models (for example, the model developed by Fair) include the assumption that the parameters of the production function are observable only when output per hour is at a peak. If firms are deferring work or temporarily enjoying increased labor effort at such peaks, then the parameters that are observed during peaks overstate the productivity that could be sustained permanently.

dence suggests that firms that hoard labor sacrifice substantial amounts of profit during downturns.[44] Because of the magnitude of foregone profits, it is important to understand why firms behave this way.[45]

Our survey has provided the first direct measurement of several variables, including the amount of hoarded labor, that are central to an understanding of the process of labor adjustment over the business cycle. Evidence regarding the reasons for and the implications of the labor adjustment patterns that we have measured should further increase our understanding of this important process.

[44] Our interpretation of the costliness of hoarding does not appear to be affected by the presence in our sample of a small group of firms that used a pure piece-rate pay system; see our working paper for details.

[45] Since 1970, after-tax corporate profits in U.S. manufacturing have averaged about 17 percent; and profits are lower during downturns. Based on this figure, the fact that 4 percent of the blue-collar labor paid for by the typical manufacturing plant during the trough quarter of a downturn was hoarded had a substantial impact on profits.

REFERENCES

Azariadis, Costas, "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, December 1975, *83*, 1183–202.

Baily, Martin Neil, (1978a) "Labor Productivity and Other Cyclical Puzzles," unpublished manuscript, December 15, 1978.

______, (1978b) "Stabilization Policy and Private Economic Behavior," *Brookings Papers on Economic Activity*, 1:1978, 11–50, and comments following by Benjamin Friedman, pp. 52–57.

Ball, R. J. and St. Cyr, E. B. A., "Short Term Employment Functions in British Manufacturing Industry," *Review of Economic Studies*, July 1966, *33*, 179–207.

Bhatia, S. L., "Capital-Labor Substitution and the Short-Run Employment Functions in the U.S. Manufacturing Sector," unpublished paper presented at the 54th Annual Conference of the Western Economic Association, June 1979.

Brechling, Frank, "The Relationship Between Output and Employment in British Manufacturing," *Review of Economic Studies*, July 1965, *32*, 187–216.

______ **and O'Brien, Peter,** "Short-Run Employment Functions in Manufacturing Industries: An International Comparison," *Review of Economics and Statistics*, August 1967, *49*, 277–87.

Chang, Julius, "An Econometric Model of the Short-Run Demand for Workers and Hours in the U.S. Auto Industry," *Journal of Econometrics*, August 1983, *22*, 301–16.

Costrell, Robert, "Overhead Labor and the Cyclical Behavior of Productivity and Real Wages," unpublished manuscript, March 1979.

Dornbusch, Rudiger and Fischer, Stanley, *Macroeconomics*, 2d ed., New York: McGraw-Hill, 1981.

Fair, Ray, *The Short-Run Demand for Workers and Hours*, Amsterdam: North-Holland, 1969.

______, "Excess Labor and the Business Cycle," *American Economic Review*, March 1985, *75*, 239–45.

Fay, Jon A. and Medoff, James L., "Labor and Output over the Business Cycle: Some Direct Evidence," Working Paper, National Bureau of Economic Research, 1985.

Freeman, Richard B. and Medoff, James L., "New Estimates of Private Sector Unionism," *Industrial and Labor Relations Review*, January 1979, *32*, 143–74.

Greer, Douglas and Rhoades, Stephen, "A Test of the Reserve Labour Hypothesis," *Economic Journal*, June 1977, *87*, 290–99.

Hall, Robert, "Employment Fluctuations and Wage Rigidity," *Brookings Papers on Economic Activity*, 1:1980, 91–141.

Hamermesh, Daniel, "Econometric Studies of Labor Demand and Their Application to Policy Analysis," *Journal of Human Resources*, Fall 1976, *11*, 507–25.

Hultgren, Thor, "Changes in Labor Cost during Cycles in Production and Business," Occasional Paper No. 74, National Bureau of Economic Research, 1960.

Ireland, N. J., Briscoe, G. and Smyth, D. J., "Specification Bias and Short-Run Returns to Labour: Some Evidence for the United Kingdom," *Review of Economics*

and Statistics, February 1973, *55*, 23–27.

Lucas, Robert, "Capacity, Overtime, and Empirical Production Functions," *American Economic Review Proceedings*, May 1970, *60*, 23–27.

Miller, Roger, "The Reserve Labour Hypothesis: Some Tests of Its Implications," *Economic Journal*, March 1971, *81*, 17–35.

Oi, Walter, "Labor as a Quasi-Fixed Factor," *Journal of Political Economy*, December 1962, *70*, 538–55.

Okun, Arthur, "Potential GNP: Its Measurement and Significance," in *1962 Proceedings of the Business and Economic Statistics Section*, 1962.

——, *Prices and Quantities: A Macroeconomic Analysis*, Washington: The Brookings Institution, 1981.

Sargent, Thomas and Wallace, Neil, "The Elasticity of Substitution and the Cyclical Behavior of Productivity, Wages and Labor's Share," *American Economic Review Proceedings*, May 1974, *64*, 257–63.

Sims, Christopher, "Output and Labor Input in Manufacturing," *Brookings Papers on Economic Activity*, 3:1974, 695–735.

Solow, Robert, "Draft of Presidential Address on the Short-Run Relation of Employment and Output," unpublished manuscript, 1964.

Tatom, John A., "The 'Problem' of Procyclical Real Wages and Productivity," *Journal of Political Economy*, April 1980, *88*, 385–94.

U.S. Department of Commerce, Bureau of the Census, *1977 Census of Manufactures*, Washington: USGPO, 1981.

An Economic Interpretation of the History of Congressional Voting in the Twentieth Century

By SAM PELTZMAN*

This paper interprets historical change in the U.S. Congress in terms of the simplest principal-agent model. I will show that profound changes in congressional voting patterns over the course of the twentieth century can be traced mainly to corresponding changes in the economic interests of their constituents. This claim may appear, at once, modest and extravagant. Modest, because the notion that agents by and large serve their principals' interests is so familiar in nonpolitical contexts. Extravagant, because economists have found the notion difficult to apply to the behavior of political agents. I begin by outlining the empirical source of this difficulty. Then I describe the main trends in twentieth-century economic history and congressional voting behavior that are the focus of subsequent empirical analysis. The analysis reveals a much closer connection between economic and political history than might be suggested by much contemporary empirical literature on the economics of voting. I conclude by attempting to reconcile these apparently divergent results.

I. The Questionable Connection Between Congressmen and their Constituents

Economists and political scientists have adduced a variety of explanations for why congressmen might rationally choose *not* to vote consistently for the interests of a majority of constituents (see, for example, Anthony Downs, 1957; James Buchanan and Gordon Tullock, 1962; George Stigler, 1971; Morris Fiorina, 1974). A more recent literature emphasizes the difficulty of linking empirically congressional voting patterns and constituent economic interests (James Kau and Paul Rubin, 1979, 1982; Joseph Kalt, 1981; Kalt and Mark Zupan, 1984; Edward Mitchell, 1979). Because my focus is also empirical, I attempt in Table 1 to provide the general reader with a sense of that difficulty.

Consider first who gains and who loses from federal tax-spending policy. This is summarized in panel A of the table. Here two measures of benefits and costs from federal programs are regressed on some state economic characteristics.[1] Each is scaled so that higher values imply more per capita "net benefits" (or lower tax rates). Neither measure is perfect,[2] but the regressions tell a similar story: the federal budget tends to redistribute wealth away from states with high incomes and large manufacturing sectors. Urbanization has a less clear-cut effect. City dwellers pay more taxes (line 2a), but perhaps those who dwell in small cities (i.e., outside SMSAs) receive net benefits (line 1b).

Panel B describes voting patterns in the Senate. The two dependent variables here, like those in panel A, are scaled so that higher values imply more support for federal taxing and spending. Again, despite their imperfections,[3] both measures tell a similar

*School of Business, University of Chicago, Chicago, IL 60637. I thank Kenneth Carl and John Markson for research assistance and anonymous referees for valuable criticism and suggestions. The financial support of the Center for the Study of the Economy and State, University of Chicago, and the Procter and Gamble Foundation is acknowledged gratefully.

[1] The particular characteristics—income, urbanization and the state's industrial mix—are chosen pragmatically. They can explain these data fairly well, and they are readily available for the much longer historical period which is this paper's main concern.

[2] Both numerator and denominator of line A.1 are estimates based on sometimes arbitrary assumptions—for example, that the burden of the deficit is proportional to taxes paid. Apart from its neglect of benefits, some of the taxes attributed to a state in line A.2 are in fact paid by residents of another.

[3] The variable *NTUA* is derived from an unweighted count of a senator's votes for increased spending or taxes on *all* roll calls dealing with taxing and spending. Thus, a vote to increase total taxes is weighted the same

(Continued, p. 658)

TABLE 1—FEDERAL TAX-SPENDING PATTERNS, SENATE VOTING PATTERNS, AND STATE ECONOMIC CHARACTERISTICS

Dependent Variable	Coefficients of[a] *HH INC* (1)	*MFG* (2)	*URB* (3)	*METRO* (4)	*SPEND/TAX* (5)	$1-TAX$ (6)	*DEMS* (7)	R^2 (8)	*SEE* (9)
A. Measures of Benefits									
1. *SPEND/TAX* × 100									
a	−11.8	−1.22	.20					.58	17.5
	(5.5)	(3.8)	(.8)						
b	−12.3	−.88	.79	−.36				.60	17.3
	(5.8)	(2.2)	(1.6)	(1.4)					
2. $1-TAX$									
a	−1.36	−.15	−.09					.64	2.66
	(4.2)	(3.2)	(2.4)						
b	−1.42	−.11	−.01	−.05				.65	2.64
	(4.4)	(1.7)	(.1)	(1.3)					
B. Pro-Spending or Pro-Liberal Voting Measures									
1. *NTUA*									
a					−.01		18.4	.24	11.4
					(.2)		(3.7)		
b						−.68	18.9	.30	10.9
						(1.8)	(3.9)		
c	.48	.03	−.25	.18			16.0	.28	11.4
	(.3)	(.1)	(.8)	(1.0)			(2.9)		
2. *ADA*									
a					−.28		26.3	.34	17.5
					(2.7)		(3.4)		
b						−2.08	30.7	.40	16.7
						(3.6)	(4.2)		
c	.65	.88	.03	−.13			29.5	.51	15.7
	(3.4)	(2.4)	(.1)	(.5)			(3.9)		
Mean	13.5	22.2	65.7	57.6	106.7	83.9	.59		
S.D.	1.6	8.2	14.6	26.1	26.1	4.3	.33		

Notes: Dependent Variables:

1. *SPEND/TAX*: estimate of federal government expenditures in a state/estimate of federal tax burden in the state. The tax "burden" includes allocation of various nonpersonal taxes (for example, corporation income taxes) to citizens of each state: average for 1975,1976,1979 × 100.

2. $(1-TAX)$: One minus ratio of Internal Revenue Service collections from individual income and payroll taxes in each state to total personal income in the state (taxes may be collected in one state from residents in another). Average for 1977–79 × 100.

3. *NTUA*: 100-average rating by National Taxpayer's Union (NTU) of senators from state for 1979–80. The *NTUA* rating is the percentage of a senator's votes that favored reduced taxes or spending, or opposed increases, so higher values of *NTUA* imply more support for taxes/expenditures. The NTU uses all votes on tax-spending issues to construct its index.

4. *ADA*: Average rating by Americans for Democratic Action (ADA) of senators from a state for 1979–80. The *ADA* rating is percentage of times a senator votes for the ADA position on a selected sample of 20 issues. The ADA counts absence or abstention as opposition; I recalculated the *ADA* rating by ignoring these nonvotes.

Independent Variables: Col. 1. *HH INC*: Median Household Income in state (thousands), 1975; col. 2. *MFG*: Percent of nonagricultural labor force in manufacturing, 1978; col. 3. *URB*: Percent of state population in urban areas, 1970; col. 4. *METRO*: Percent of state population in standard metropolitan statistical areas (SMSAs), 1978; cols. 5,6: see lines A.1,A.2; col. 7. *DEMS*: Number of Democrat senators from state/2, 1979–80.

Sources: *SPEND/TAX*, $(1-TAX)$, and independent variables: *Statistical Abstract of the United States*; *NTUA* and *ADA* from data supplied by NTU and ADA, respectively.

[a] The *t*-ratios are shown in parentheses below coefficients.

story: apart from the tendency for Democrats to vote for more spending/taxes (col. 7),[4] there is either no connection or a *perverse* connection between the interests of constituents and the votes of their senators. For example, holding party constant, pro-tax/spending voting is either uncorrelated (line 1a) or *negatively* correlated (1b, 2a, 2b) with the direction of benefits from those policies. Further, the characteristics most clearly *negatively* correlated with net spending benefits—income and manufacturing—are either uncorrelated (line 1c) or *positively* correlated (2c) with voting for larger federal spending.

The results in Table 1 are only a suggestive introduction to the empirical literature on congressional voting.[5] But I believe that the reader will find Table 1 consistent with an important broad conclusion of that literature, namely that there is a large "inertial" component (often labelled "ideology") in congressional voting: "liberal" or "conservative" voting patterns tend to persist from issue to issue, whether or not they seem consistent with constituent interests.

My task here will be to see if this sort of inertia is evident in a much longer historical perspective: is it just a recent anomaly or a main feature of the history of legislation? Have long-period changes in economic circumstances had any substantial connection with historical changes in congressional voting patterns? If so, is the connection consistent with changes in economic interests? In answering these questions, I shall try to disentangle the economic analysis of legislation from some of the idiosyncracy of American history. A clear regional pattern underlies data such as those in Table 1: southern congressmen tend to be more conservative than northerners (Kau and Rubin, 1979), while per capita income in the South is also lower than the North. If "economic interest" is supposed to be *all* that matters, this positive correlation between income and liberalism appears perverse. But it can be consonant with a world in which economic interest is either *important* or *unimportant*. There is a long American history of regional political division, as well as of regional economic differences, and in any cross-section correlation between, income and liberal voting will reflect a mix of historical and economic forces. Accordingly, we cannot tell from the positive correlation we observe today whether economic interest and other factors (like the legacy of the Civil War) tug strongly in opposing directions, or whether economic interest matters little next to "history," or whether history matters little and political redistribution is merely a "normal" consumer good "bought" most heavily by rich constituencies. A major goal herein is to sort these possibilities out.

The paper is organized as follows. Section II sets out the basic economic and political history that underlies the subsequent empirical analysis. Section III describes how I measure both economic interest and voting behavior and what I think these measures tell us. Sections IV and V use these measures in empirical analyses designed to gauge the importance of economic forces in shaping the history of congressional voting. The main findings of this analysis are that: 1) voting patterns in a typical Congress reveal an interplay of economic forces and often con-

as a vote to increase the budget of the Battle Monuments Commission. The variable *ADA* is derived from votes for the Americans for Democratic Action (ADA) position on a selected sample of 20 issues deemed "important" by that organization. These issues are not limited to tax-spending matters, as with *NTUA*. However, they will usually include the more important tax-spending issues in a Congress, since the ADA has traditionally favored expansion of federal spending, especially on domestic programs.

[4] My 1984 article shows that in popular elections, Democrats tend to draw votes from lower-income voters. Thus, since federal tax-spending policy appears "progressive," the tendency for Democrats to be pro-spending/taxing is consistent with a simple principal-agent story.

[5] The regressions in panel A do not, for example, capture the within-state variance of benefits or costs. For example, if taxes are sufficiently concentrated within a state, a high tax-income ratio or low spending-tax ratio for a state may still leave a majority of residents with net benefits from the federal budget. Also, note that even if the *total* impact of the federal budget is progressive, this need not be true of the *additions* to the budget at issue in congressional votes in 1980. Hence a positive correlation between "pro-spending" votes and income is not necessarily "perverse" from the standpoint of the economic interests of constituents.

flicting regional cultural preferences, but these preferences seem remarkably stable over time. More importantly: 2) this "historical inertia" has not prevented a profound change in voting patterns over the course of the twentieth century; and 3) this change seems entirely attributable to a corresponding change in economic interests in redistributive legislation. Section VI reexamines some puzzles such as those in Table 1 in light of these findings.

II. The Historical Background

One fact dominates the twentieth-century American economic history which is relevant to this paper: states and regions have become economically more homogeneous. Figure 1 illustrates this for three characteristics.[6] The solid lines track dispersion across states. They begin turning down c.1920 (over a period centered on 1920) or earlier, and by 1980, have fallen by about 25 percent (urbanization) to about 70 percent (income) from their peak values. The same movements are evident when states are grouped into the nine census regions (broken lines). The generally small vertical distances between these roughly parallel lines imply that the homogenization process has been mainly a regional phenomenon.

Any economic explanation of congressional voting has to come to grips with this profound narrowing of regional economic differences and, presumably, of economic interests. However, if this would seem to imply a similar narrowing of political differences, the crude data belie such similarity. Figure 2 portrays the history of a few measures of state and regional political difference. The top panels focus on party membership in Congress. Any detectable narrowing of sectional differences in party membership seems confined to the most recent two decades or so, or thirty years after the onset of the economic homogenization. In addition, there is much more intraregional variability in party membership than in the economic data.

[6] See the notes to Table 3 for explanation and sources of these variables.

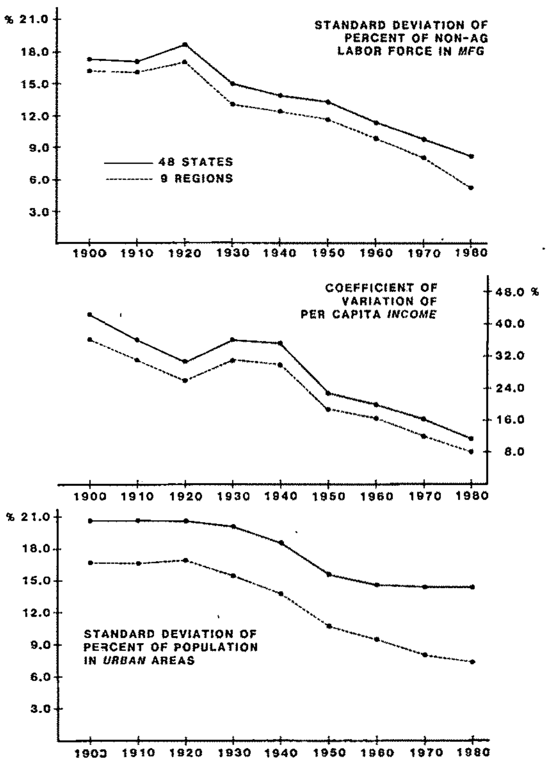


FIGURE 1

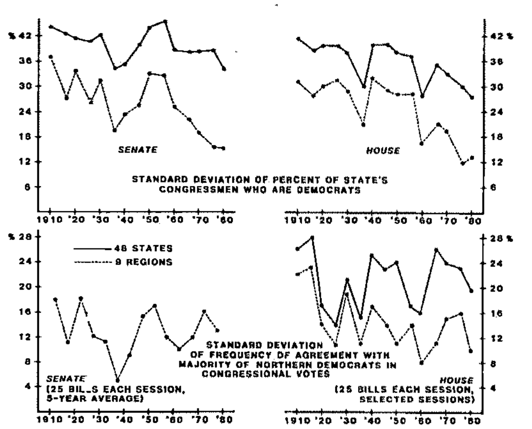


FIGURE 2

Because the heterogeneity of American parties may make party membership too crude a measure of political difference, the bottom panels of the figure focus directly on voting behavior. The specific measure underlying

these panels is the average frequency with which a state's or region's congressmen voted with the northern Democrat majority on successive samples of 25 votes taken in a session of each House. (The sample and the voting variable are described more fully in the next section.) The bottom panels of Figure 2 show time-series of the standard deviation of these frequencies across states and regions. They appear trendless. As with party membership, agreement between these political data and the economic data in Figure 1 escapes the naked eye.[7]

I will attempt to bridge this apparent disjunction between economic and political differences in Section VI.

III. The Calibration of Economic Interest and Congressional Voting

Are the data discussed in the two previous sections hiding some historically durable connection between the way congressmen vote and their constituencies' economic stake in that vote? I want to answer this with conventional, easily replicable statistical analysis, and this demands plausible empirical summaries of both the votes and the economic interests. This is a formidable demand. The dramatic growth of the size and scope of the federal government implies that the menu of issues facing Congress and the nature of the stakes may be much different today than in 1900. My first task, therefore, is to show that there is enough historical consistency in both the nature of the "stakes" and the pattern of voting to motivate the reader's interest in the subsequent empirical analysis.

To get at the consistency of the stakes, I focus on the redistributive element in federal policies. I showed earlier that contemporary budget policy redistributes wealth away from high-income, manufacturing-intensive, and, possibly, large urban areas. Unfortunately, a long history of the redistributive effects of federal spending and taxes is unavailable. However, the income-progressivity feature of the redistribution seems to go back at least to 1950 (Morgan Reynolds and Eugene Smolensky, 1977). Moreover, there are enough income tax. collection data for replication of the sort of regression on line A.2a in Table 1 for most of the relevant historical period and they also show a fairly consistent pattern. I regressed the log of the income taxes/income ratio in each state in the years 1920, 1930, ..., 1970 on the same three variables as in Table 1 plus time dummies and obtained:

$$(1)\quad \mathrm{Log}(Tax/Income)_{it}$$

$$= \text{Constant} + \text{coefficients} \times YEAR \text{ dummies} + \underset{(5.0)}{.59}\left(\frac{Per\ Capita\ Income_{it}}{US\ Per\ Capita\ Income_t}\right) + \underset{(5.4)}{.011}\, URB_{it} + \underset{(3.8)}{.006}\, MFG_{it}$$

$$R^2 = .95,\ SEE = .32,$$

where i = state, $t = 1920, 1930, \ldots, 1970$.[8]

When the regression is estimated separately for each of the six years, the coefficients are not always significant, but 17 of the 18 are positive and all 18 simple correlations are significantly positive. Thus, the few available pieces of data give at least a broad hint about the nature of the stakes in political redistribution: the interests of constituents from high-income, manufacturing-intensive, and urban areas have *generally* been asking for opposition by their congressmen to expansion of the federal budget for

[7] We cannot, of course, rule out the possibility that increased homogeneity has affected the content of proposed legislation. For example, rural congressmen will oppose either a 1 percent tax or a 10 percent tax on farm incomes, but the former clearly entails a smaller difference between urban and rural interests.

[8] The sample is the 48 continental states, except Delaware, which is an extreme outlier in the earlier years. Tax data are from U.S. Internal Revenue Service and comprise personal income and employment taxes. See Table 1 for definition of *URB* and *MFG*, and see Table 3 for sources. The t-ratios are shown in parentheses.

at least the last sixty years.[9] Accordingly, in the subsequent analysis I use the income, urbanization, and manufacturing measures to summarize the diversity of interest in redistribution across congressional constituencies.

I will compare these interests with an equally simple summary measure of congressional voting behavior: the extent of support for the position taken by the majority of northern Democrats. While the details of my treatment of the Senate and House differ, nothing essential is lost if I first describe my procedure for senators and then explain why I think it results in an historically consistent measure of voting behavior.

From each Congress, from the 63rd (when popularly elected senators began sitting) to the 96th (1980), I drew a sample of 25 bills on which record votes were taken.[10] I included only votes on "economic" (i.e., nondefense budgetary and economic regulatory) issues where the winning side had less than a 2–1 margin and where over half the senators voted.[11] These criteria were meant to limit the analysis to controversial issues with potentially significant redistributive elements. Each senator's vote on each bill was then coded +1 if he voted in favor of the position taken by the majority of northern Democrats, or 0 if he voted the other way. Call the dichotomous variable liberal (*LIB*). I then extracted the *regional*[12] elements of *LIB* on each vote from the regression

$$(2) \qquad \left[LIB_{ij} - \overline{LIB}_j\right] = \sum_k B_{kj} \cdot D_{ik} + \text{error term},$$

where LIB_{ij} = ith senator's vote on the jth bill, $\overline{LIB}_j$ = average value of LIB on the jth bill, ($i=1,\ldots,96$ senators; $j=1,\ldots,25$ bills in each session of Congress), D_{ik} = dummy variable = +1 if senator i represents a state in census region k; 0 otherwise; $k=1,\ldots,9$ census regions, B_{kj} = bill-specific regression coefficient of D_{ik}.

For each bill, I also estimated a regression with the same left-hand side variable as (2) and right-hand variables:

$$(3) \qquad \text{regional dummies} + C_j\left(P_i - \overline{P}_j\right)$$

where $P_i = +1$ if senator i is a Democrat,[13] 0 if Republican; $\overline{P}_j$ = average value of P on bill j; C_j = bill-specific regression coefficient.

An important motivation for holding party effects constant, as in (3), is to control for

[9] The qualifications to this conclusion would include: 1) Non-income taxes—customs and excises—were far more important revenue sources prior to World War II than today, and there is no presumption that their geographical distribution followed that of income taxes. In addition, customs duties provided net *benefits* to protected industries. 2) Other forms of federal government activity (for example, regulation) could have different distributional implications than spending and taxes, so general opposition to expansion of government economic activity need not be in the interest of, for example, a high-income constituency. 3) One could surely lengthen the list of variables potentially related to the interest in redistribution—I intend my short list as a plausible summary, not an exhaustive summary.

[10] The voting data are from tape files compiled by the University of Michigan Inter-University Consortium for Political and Social Research (ICPSR). The files give a brief narrative description of each record vote and the position taken by each congressman.

[11] Today a record vote almost always attracts substantial participation, but prior to the 1930's it was not uncommon for a majority of Senators to be absent without declaring a preference on a vote. I counted any expression of support or opposition—paired votes and announced positions—as a "vote."

[12] Since the economic data are available by state, my goal of relating political to economic behavior suggests extracting the state, rather than regional, regularities in *LIB*. However, with no more than two senators from a state voting on a bill, the "standard error" of the "state" regularity is rather high. Also, the dominant role of regional differences in the interstate variance of economic variables suggests that "region" is a sensible level of aggregation for my purpose.

[13] Senators who belonged to neither major party were assigned to one of the two main parties as follows. For each of the 25 bills in each session, I calculated the average value of *LIB* for Democrats and for Republicans. Then I correlated the third-party senator's *LIB* with the difference between the Democrat and Republican average. If the correlation was significantly positive (negative), I called the senator a Democrat (Republican). In virtually every case these correlations were so high (on the order of ±.8) that there was little doubt about where to put the ostensibly maverick senator.

some of the effects of the intraconstituency diversity of economic interest.[14]

The measure (denoted B_{kt}) of a region's liberal voting tendency for a session which I use in subsequent analyses is the average of each region's B_{kj} across the 25 sample votes in that session. In substance, each B_{kt} is just a count of the relative frequency of agreement with the northern Democrat majority. For example, suppose that the typical New England senator voted *LIB* 30 percent more frequently than the average senator and 20 percent more frequently than the average of his party colleagues in 1920. Then B_{kt} from (2) = +.3, while B_{kt} from (3) = +.2 for t = 1920. Since the average of the B_{kt} across regions is roughly zero[15] in every session, any B_{kt} measures the liberalism (*LIB*) of a region's senators *relative* to all senators in that session.

My *LIB* measure and its derivatives are replicable, but are they measuring any historically consistent political behavior? I want to suggest that they measure just what the *LIB* acronym might suggest: a propensity to support progressive redistribution. This is not a controversial suggestion for the period since the New Deal, since Democrats, especially in the North, have been in the vanguard of support for expansion of the progressive tax-spending structure. But I want to argue that the same interpretation is plausible for the pre-New Deal period as well. In my sample of bills from the pre-New Deal Congresses, I found two issues recurring often enough to permit meaningful generalization. These were bills to change 1) the level or degree of progressivity of federal income tax rates, and 2) tariff rates on imports of manufactured goods. For each such bill in my sample, I computed the difference between the percentage of northern Democrat and Republican votes in favor of increased taxes or progressivity or of increased tariffs on manufactured goods. The results, summarized in Table 2, indicate the clear preference of northern Democrats for higher/more-progressive income taxes and lower tariffs on manufactured goods. The distributive implications of the northern Democrats' income tax policy are reasonably clear. I am unaware of any evidence on the incidence of tariffs in this period. However, the first-order protective effects of tariffs in the early twentieth century would appear to be regressive: they raised incomes of individuals (owners and workers in manufacturing) who, in a society with a large, low-income agricultural sector, had above-average income. So, with some uncertainty, it seems plausible to characterize the economic policy favored by northern Democrats as consistently pro-redistributive for all of this century.

The main conclusions from this brief tour of twentieth-century economic and political data are: 1) high values of B (or *LIB*) con-

TABLE 2—DIFFERENCES BETWEEN NORTHERN DEMOCRAT AND REPUBLICAN POSITIONS ON INCOME TAX AND TARIFF ISSUES, VOTE SAMPLES FOR 61ST–74TH CONGRESSES

Issue and Position	Number of Votes: Senate	Number of Votes: House
A. For Higher Income Taxes or More Progressivity		
1. Total Bills	21	11
a. % Northern Dems (*ND*) For > % Republicans (*GOP*)	14	10
b. % *ND* < % *GOP*	3	0
c. No significant difference	4	1
B. For Higher Mfd. Good Tariffs:		
1. Total Bills	32	21
a. % *ND* > % *GOP*	0	0
b. % *ND* < % *GOP*	30	20
c. No significant difference	2	1

Source: ICPSR File.

[14] See fn. 4. More generally, a senator's constituency is not his "state," but those parts of the electorate relevant to putting him in office and keeping him there. See Richard Fenno (1978) for a discussion of this point. My earlier article shows that the economic characteristics of this "supporting electorate" differ systematically according to a senator's party. Thus, for example, the typical constituency of Democrats from rich states gains less from redistribution than that of Democrats from poor states, but not necessarily less than that of Republicans from poor states.

[15] More precisely, since the number of states per region varies, a state-weighted average of B_{kt} across regions is zero. At the same time, my procedure induces a purely statistical bias in favor of positive B's for the North and negative B's for the South. The nature and size of the bias is discussed in fn. 18 below.

note congressional support for a progressive redistributive economic policy; 2) the main beneficiaries of such policies should be found in poor, non-urban, non-manufacturing-intensive areas.

IV. Empirical Results: The Historical Connection Between Economics and Politics

In light of previous discussion, I analyze the link between economics and politics in terms of three models of the process:

1. Only economic forces matter in politics.
2. Economic forces don't matter at all, that is, historical regional political differences persist in spite of economic change.
3. Both economics and history matter.

Operationally, this entails estimating regressions which are variants of the general form

$$B_{kt} = K + \Sigma M_i \cdot ECON_{ikt} + \Sigma N_k \cdot R_k, \tag{4}$$

where B_{kt} = the relative *LIB* of region (or state) k's congressmen in year t; $ECON_{ikt}$ = proxies for k's economic (*ECON*) interest in liberal (*LIB*) votes in t; R_k = a set of time-invariant regional (or state) dummies; K, M, N = parameters.

The overall fit and the accuracy of the parameters of regressions on various subsets of the *ECON* and *R* variables will be used to draw inferences about which of the three models best describes the voting behavior summarized by the B_{kt}.

For the Senate, I have estimates of each region's B_k for each session of Congress from the 63rd through the 96th (1913–80). The *ECON* variables suggested by the preceding discussion (income, urbanization, manufacturing intensity) are consistently available only every decade. As a compromise between the rich political data and lean economic data, I computed (a) 5-year averages of the B_k and (b) semidecadal values of the economic variables (by interpolation of the decade-end values). The economic variables, like the political variables, are each measured relative to their period means. This yields 126 observations on relative political behavior and relative economic conditions: one for each of nine regions in each of 14 periods ending 1915,1920,...,1980. Table 3 summarizes the results of implementing the three models with this body of data. There are two dependent variables: relative liberalism of a region's senators without regard to party in columns 1–3 and adjusted for party (i.e., the average deviation of a region's Democrats and Republicans from their party's average) in columns 4–6. Each triplet implements the three models successively. The clear result is that the eclectic model which includes both *ECON* variables and regional dummies yields the most adequate description of the voting history. The eclectic model has greater explanatory power (as measured by *SEE*), and coefficients of both the economic and regional variables tend to be estimated more precisely than in the special-purpose models. Thus, these data suggest that economic forces have combined with persistent regional differences to produce the observed Senate voting patterns.

The same conclusion emerges from analysis of voting in the House of Representatives. Here, to economize on computation costs, I sampled votes only from sessions ending in year 0 and year 6 of each decade from 1910 through 1980. However, with over 400 representatives, there are sufficient degrees of freedom to allow the state, as well as the region, to be the unit of analysis. Accordingly for each House vote,[16] I estimated equations like (2) and (3) except that 48 state dummy variables, rather than just 9 regional dummies appear in them. Then, as with the Senate, I averaged the coefficients of the dummies over the 25-vote regressions in each session to provide my measure of the liberalism of a state's House delegation for that session. This yielded a sample of 720 observations on the measure of political behavior—one for each of 48 states in each of 15 sessions ending 1910,1916,1920,1926...1980. The results of implementing the same three models as for the Senate are shown in Table 4. Again, the

[16]As with the Senate, I drew samples of 25 votes—on criteria like those for the Senate vote samples—from each of the sessions included in the analysis.

TABLE 3—REGRESSIONS OF POLITICAL LIBERALISM IN THE SENATE ON ECONOMIC CHARACTERISTICS AND REGIONAL DUMMIES; 5-YEAR PERIODS: 1910–15 TO 1975–80, 9 REGIONS

	Dependent Variable and Model, Liberalism (B_{kt})					
	No Adjustment for Party			Net of Party Effect		
Independent Variables:	*ECON* (1)	Region (2)	Both (3)	*ECON* (4)	Region (5)	Both (6)
Economic:						
1. *MFG*	−.281		−.576	−.199		−.436
	(2.2)		(1.5)	(1.9)		(1.8)
2. *PCI*	−.319		−.501	.018		−.374
	(3.3)		(5.1)	(.2)		(5.9)
3. *URB*	.288		−.467	.232		−.195
	(1.3)		(1.5)	(1.3)		(1.0)
Regional Dummies:						
4. New England (NE)		−.071	.138		.029	.171
(ME, NH, VT, MA, RI, CT)		(2.2)	(2.9)		(1.4)	(5.6)
5. Mid-Atlantic (MA)		−.060	.227		.037	.223
(NY, NJ, PA, DE, MD)		(1.9)	(5.9)		(1.8)	(8.9)
6. EN Central (ENC)		−.022	1.80		.032	.166
(OH, IN, IL, MI, WI)		(.7)	(4.8)		(1.5)	(6.9)
7. WN Central (WNC)		−.064	−.193		.040	−.042
(MN, IA, MO, ND, SD, NE, KS)		(2.0)	(5.6)		(1.9)	(1.9)
8. S Atlantic (SA)		.039	−.128		−.103	−.207
(VA, WV, NC, SC, GA, FL)		(1.2)	(2.4)		(4.9)	(6.0)
9. ES Central (ESC)		.074	−.210		−.037	−.222
(KY, TN, AL, MS)		(2.3)	(3.7)		(1.8)	(6.1)
10. WS Central (WSC)		.047	−.148		−.084	−.219
(AR, LA, OK, TX)		(1.5)	(4.2)		(4.0)	(9.6)
11. Mountain (MT)		.007	−.077		−.002	−.060
(MT, ID, WY, CO, NM, AZ, UT, NV)		(.2)	(1.3)		(0.1)	(1.5)
12. Pacific (PAC)		.050	.209		.085	.188
(WA, OR, CA)		(1.6)	(4.9)		(4.1)	(6.9)
R^2	.24	.17	.57	.09	.38	.69
SEE	.112	.119	.087	.093	.078	.057

Notes: Dependent Variables: For cols. 1–3, the variable is derived from regression estimates of equation (2) (see text). Coefficients of the regional dummies in those vote regressions are averaged over the 25 votes in each session. Then these regional averages are further averaged over 5-year periods as follows. Since each session of Congress ends in an even-numbered year, let $\overline{B}_{k2}, \overline{B}_{k4}, \ldots, \overline{B}_{k10}$ represent the average coefficient for region k in the 25 vote regressions for a session ending in year 2, year 4, etc., of a decade. The dependent variable for the first 5-year period in a decade is $1/2.5(\overline{B}_{k2} + \overline{B}_{k4} + .5B_{k6})$ and that for the second 5-year period is $1/2.5(.5\overline{B}_{k6} + \overline{B}_{k8} + \overline{B}_{k10})$. (For the period ending 1915, the calculation is $(1/1.5)$ $(\overline{B}_{k4} + .5B_{k6})$, because I exclude data from Senates without popularly elected senators.) The variable for cols. 4–6 is constructed in the same way except that it is based on regression coefficients from equation (3) which includes a party dummy. The vote data on which these variables are based come from the ICPRS files.

Independent Variables: Economic—each of these has the form $(X_{kt} - \overline{X}_t)$ where X_{kt} is an average over the states in a region in year t and $\overline{X}_t$ is a 48-state average for t (Alaska and Hawaii are excluded), and t is every 5 years from 1915 through 1980. The definition and sources of the state data are

1. *MFG* = percent of nonagricultural labor force in manufacturing. For 1940, see the U.S. Bureau of Labor Statistics, *Handbook of Labor Statistics*; for 1910–40, I use (gainful workers in manufacturing/all nonfarm gainful workers) from Everett S. Lee et al. (1957). I regressed the BLS data on the Lee et al. data for 1940 and 1950 ($R^2 > .95$) and used the regression coefficients and the Lee et al. data to generate estimates of the BLS measure for years prior to 1940.

2. *PCI* = Log of per capita personal income. For 1930–80, from U.S. Bureau of the Census, *Historical Statistics of the United States* and *Statistical Abstract...*; for 1920, Maurice Leven (1925); for 1910, Lee et al.

3. *URB* = Percent of state population in urban areas from *Historical Statistics*...and *Statistical Abstract*....

For the middle year of each decade, each variable is a linear interpolation of the value at the beginning and end of the decade. Regional dummies each = +1 if the dependent variable is for region k, 0 otherwise. The regions are as defined by the Census Bureau, except that, as in Lee et al., Delaware and Maryland are moved from the South

TABLE 4—REGRESSIONS OF HOUSE VOTING PATTERNS ON ECONOMIC CHARACTERISTICS AND STATE DUMMIES, 15 CONGRESSES (1910–80), 48 STATES

	Dependent Variable and Model Liberalism (B_{kt})					
	No Adjustment for Party			Net of Party Effect		
Independent Variables	*ECON* (1)	State (2)	Both (3)	*ECON* (4)	State (5)	Both (6)
Economic:						
1. *MFG*	−.134		−.273	−.042		−.151
	2.1		1.7	1.1		1.5
2. *PCI*	−.323		−.327	.093		−.163
	7.5		5.4	3.5		4.4
3. *URB*	.234		−.878	.070		−.483
	3.2		6.6	−1.6		6.0
State Dummies: Regional Averages						
4. New England		−.085	.104		.042	.144
		(.12)	(.35)		(.02)	(.12)
5. Mid-Atlantic		−.002	.297		.042	.201
		(.08)	(.13)		(.01)	(.07)
6. EN Central		−.060	.124		.026	.125
		(.05)	(.09)		(.02)	(.05)
7. WN Central		−.087	−.213		.058	−.010
		(.12)	(.22)		(.05)	(.06)
8. S Atlantic		.100	−.082		.093	−.189
		(.04)	(.09)		(.06)	(.06)
9. ES Central		.115	−.144		−.057	−.193
		(.05)	(.11)		(.09)	(.12)
10. WS Central		.150	−.006		−.066	−.148
		(.02)	(.09)		(.03)	(.04)
11. Mountain		.025	.081		−.017	−.048
		(.11)	(.16)		(.04)	(.06)
12. Pacific		−.011	.168		.059	.154
		(.04)	(.14)		(.02)	(.06)
R^2	.09	.30	.45	.02	.31	.43
SEE	.205	.185	.165	.126	.109	.099

Notes: See Table 3 and text for definitions and sources of variables. Each of the 720 observations is on a state in a session. The sessions are at 4- or 6-year intervals 1910,1916,1920,1926...1980. The dependent variable is the average coefficient on a state dummy variable from 25 vote regressions like equations (2) and (3) estimated in each session. The regressions used for estimating the dependent variable for cols. 4–6 include a party dummy. The Economic variables are deviations of state variables from a 48-state average for each session. (District-level data are unavailable for the whole period.) The regressions in cols. 2, 3, 5, and 6 include 48 state dummies. Their coefficients are summarized here by region: the standard deviations of the state coefficients in a region are shown in parentheses below the regional mean of these coefficients. These means have the same interpretation as their counterparts in Table 3—i.e., they show the difference between frequency of liberal voting in a region and in whole House (cols. 2,3) or within a party (cols. 5,6). I reestimated the regressions using weighted least squares, with (number of congressmen)$^{-1/2}$ as the weight, because analysis of the residuals revealed some heteroscedasticity. But the coefficients and t-ratios of the economic variables were virtually identical to those reported here.

Atlantic to Middle Atlantic region. All variables except dummies are in fractions of 100, for example, the coefficient of *MFG* in col. 1 means: "in a region where *MFG* is 10 percent above the national average, senators will vote *LIB* 2.81 percent less frequently than the average senator." The coefficients of the regional variables in col. 3 are given for *MFG* = *PCI* = *URB* = 0, for example, the coefficient for New England in col. 3 means "if New England had the national-average economic characteristics, its senators would have voted *LIB* 13.8 percent more frequently than the average senator." For cols. (4)–(6), these statements apply to the deviation of the average senator in a region from the average of all his party colleagues. The residuals from the regressions in cols. 3 and 6 are both serially correlated (r = .35). (This is partly induced by the averaging process used in generating the data, in that the dependent variable for adjacent periods shares the common term $.2\bar{B}_{k6}$ (see above).) Accordingly, I reestimated these regressions via *GLS*. The results were virtually identical to those reported here, except that t-ratios were smaller than those reported here (shown below the coefficients in parentheses). For example, the *GLS* t-ratios for *MFG*, *PCI*, and *URB* were 1.1, 4.2, and 1.4 for the col. 3 regression, and 1.0, 4.6, and 1.1 for col. 6.

eclectic models (cols. 3 and 6) fit the data significantly better than either the "economics only" or "history only" models.[17]

It is, at this point, only convenient shorthand to describe the preceding results as showing that economic change modifies historical inertia. These results are conceptually similar to those of the previously cited literature on contemporary voting in that "noneconomic" variables—here regional dummies—have important marginal explanatory power. So, one could allude to regional differences in ideology as easily as to "historical inertia" (or "tastes" or "unmeasured variables"). Later, I provide some motivation for my shorthand by showing that there is in fact considerable inertia in the history—that is, that the coefficients of the regional dummy variables in cols. 3 and 6 of Tables 3 and 4 are stable. But, however they are labeled, the statistical significance and large magnitudes of these regional differences are a challenge to future research:[18] can we find measurable regional characteristics with histories much different from those already in my analysis whose inclusion would reduce the explanatory power of the regional dummies?[19]

In the next section, I use the results in Tables 3 and 4 to analyze historical *changes* in voting patterns, but they also reveal interesting regularities in the average "levels" of political behavior:

1. The coefficients of the economic variables in columns 3 and 6 are "sensible"

[17]The economic variables appear to have less marginal explanatory power over the "location only" model for the Senate (compare the change in *SEE* from cols. 2 to 3 or cols. 5 to 6 in the two tables). However, this is due to the larger number of locational dummies in the House regressions, and the consequent ability to better "explain" state idiosyncracies. If the House data are, like the Senate data, grouped into regions and the Table 3 regressions replicated exactly (i.e., with regional economic and dummy variables), the *SEE*'s are as follows:

col. 1: .116 col. 4: .086 col. 2: .124

col. 5: .075 col. 3: .087 col. 6: .057.

Here the gain in *SEE* from col. 2 to 3 or col. 5 to 6 is around twice that in Table 4 and almost identical to that in Table 3 for the Senate.

[18]A very small part of these differences is due to the statistical bias alluded to in fn. 12. The bias arises because my *LIB* measure is, in part, regionally based. Since northern Democrats are a subset of northern senators, measures of how the average northern senator and the northern Democrat majority vote will tend to agree even if there are no substantive regional differences. Thus, the coefficients of the northern regional dummies will tend to be positive. However, the *magnitude* of this bias is much too small to account for the regional differences observed in cols. 3 and 6 of Tables 3 and 4.

To isolate the magnitude of the bias, assume that every one of 100 senators votes randomly on every bill, so there are no substantive regional differences at all. My procedure would then select the X Democrats among the 70 (in round numbers) northern senators and use their randomly generated majority position to define a *LIB* vote. The random process generating this majority produces a mean of $.5X + .399\sqrt{X}$ votes in favor of that position among northern Democrats (under a normal approximation). This implies the following differences between the mean probability ($\bar{P}$) of a vote favorable to the northern Democrat majority and the corresponding regional probability;

$$P_{NORTH} - \bar{P} = .399\sqrt{X}\,((1/70) - (1/100))$$

$$P_{SOUTH} - \bar{P} = -.399\sqrt{X}/100.$$

These differences (or, equivalently, regression coefficients on regional dummies) are as follows for various values of X which span the range of twentieth-century political experience:

X	$P_{NORTH} - \bar{P}$	$P_{SOUTH} - \bar{P}$
10	.005	−.013
30	.009	−.022
50	.012	−.028
60	.013	−.031

Thus, statistical bias implies coefficients of regional dummies which average only one-tenth or so of those in cols. 3 and 6 of Tables 3 and 4.

[19]My earlier paper argues that differences among senators in sources of electoral and financial support can explain much of the apparent ideological inertia in contemporary voting. For example, senators from the same state will vote differently because they drew votes and funds from systematically different groups within the state. This finding suggests that we look to persistent regional differences in these sources of electoral and financial support for an economic explanation of historical inertia—for example, senators from historically conservative regions may have drawn support from the upper end of their state's income distribution. On this argument, the low voting participation of low-income blacks in the South for most of the twentieth century may help explain that region's historical conservatism. If so, the size of the regional coefficients would be reduced if we could substitute something like income per voter for income per capita in the regressions.

(unlike their contemporary counterparts in Table 1): these variables, all of which we have seen to be negatively correlated with benefits from redistribution, are also negatively correlated with voting for redistribution.

2. Economics and history have tended to be opposing forces. This is revealed by comparing the regional coefficients in columns 2 and 5 (of either Table 3 or 4) with their counterparts in columns 3 and 6. The latter isolate the effect of history, because economics is separately accounted for in these regressions. The coefficients in columns 2 and 5 show the *net* impact of history and some average of economic forces. Note that these "net" measures range less broadly ($\pm.10$, very roughly) than the "pure history" measures (the range in cols. 3 and 6 is around $\pm.20$). Thus, economic forces have typically dampened the effects of history. In particular, they have dampened the South's conservatism and the North's liberalism.[20]

3. Economic forces affect both the behavior of congressmen from the same party and the party composition of Congress. To see this, compare the coefficients of the economic variables in column 6 (which describe intraparty behavior) with those in column 3 (where interparty differences are not removed). The former have the same signs but are smaller absolutely than the latter. This says that, for example, higher income in a region makes both Democrats and Republicans in that region more conservative (col. 6, line 2) but it makes the average congressman still more conservative (the absolute value of col. 3, line 2 exceeds that of col. 6, line 2 in both tables). This implies an increase in the number of more conservative Republicans representing that region.[21] Apparently congressmen are at least partly constrained by the central position of their party: they move away from it to accommodate the economic interests of their constituency but not always far enough to remain in office. This finding helps explain why narrowing of interregional differences in party composition (for example, the breakup of the "Solid South") lags behind the narrowing of economic differences: the initial changes in economic interest can be accommodated by shifts in position within a party, but their cumulation over time eventually breaks a party's hold on a region.

V. The Economic Basis of Historical Change in Congressional Politics

Thus far I have shown that voting patterns in a typical Congress can be described by the interaction of economic forces and persisting regional differences, rather than by the working of economic forces alone. In this section, I examine critically the logical corollary of that description—that the only source of historical *change* in voting patterns is economic change. How well does this corollary describe the changes that have occurred over the course of the twentieth century? Are the political changes attributable to economic change substantial or trivial? Are they substantial enough to overcome or just slightly modify the otherwise persisting regional differences?

[20] The same dampening also shows up within regions. In Table 4, note the smaller intraregional variation (the entries in parentheses) of the net coefficients in cols. 2 and 5 vs. their pure history counterparts in cols. 3 and 6. Yet another indication of the opposition of economics and history is revealed by subtracting the coefficients in col. 3 (or col. 6) from their counterparts in col. 2 (or col. 5). This operation shows the average direction of the effect of regional economic forces in modifying history, and it almost always yields a number opposite in sign to the impact of history. For example, line 4, col. 6 of Table 3 tells us that, *holding economic forces constant*, the typical New England senator votes liberal 17 percent more frequently than his party colleagues. But when economics is not held constant (col. 5), this excess liberal frequency is only 3 percent. The implication is that economics has, on average over the twentieth century, pulled against these senators' "natural" liberalism.

[21] Compare also the sizes of the differences between the regional coefficients in cols. 2 and 3 on the one hand, with those between cols. 5 and 6 on the other. These measures of the impact of economic forces in modifying history tend to be smaller absolutely within parties (col. 5–col. 6) than within the Senate or House as a whole (col. 2–col. 3), though they go in the same direction. Again, the implication is that if economics impels toward, for example, more conservative voting, part of the impulse is reflected in the replacement of members of the more liberal party.

TABLE 5—CHANGE IN FREQUENCY OF LIBERAL VOTING IN THE SENATE, c.1920–c.1975, REGIONS

Liberal Voting Measure	Region and Voting Measure $\times 10^3$									SD Across Regions
	NE (1)	MA (2)	ENC (3)	WNC (4)	SA (5)	ESC (6)	WSC (7)	MT (8)	PAC (9)	(10)
A. Within Senate (No Party Adjustment)										
1. Actual c.1920	−225	−118	−75	−71	206	188	171	0	−58	152
2. Actual c.1975	154	89	148	−6	−136	−139	−119	−90	135	128
3. Change (2.−1.)	379	207	223	65	−342	−327	−291	−90	193	273
(Col. 3, Table 3)										(r = .95)
4. Predicted Change	290	189	109	−121	−214	−274	−230	−63	115	204
5. Residual (3.−4.)	89	18	114	186	−128	−53	−61	−27	78	100
B. Within Parties (Net of Party Effect)										
1. Actual c.1920	−44	−30	18	50	−12	−22	−51	−13	48	37
2. Actual c.1975	151	212	107	25	−198	−175	−167	−64	141	159
3. Change (2.−1.)	195	242	89	−25	−186	−153	−116	−51	93	153
(Col. 6, Table 3)										(r = .96)
4. Predicted Change	182	121	66	−85	−147	−192	−148	−26	79	134
5. Residual (3.−4.)	13	121	23	60	−39	39	−13	−25	14	49

Note: The entries in the table are based on the measure of liberal voting analyzed in Table 3—i.e., the frequency of agreement with the majority position of northern Democrats, for examples, the −225 on line A.1, col. 1, means that New England senators voted with the northern Democrat majority 22.5 percent less frequently than the average senator (over a period centered on 1920). The entries on lines A.1 and A.2 are the average of the dependent variable in Table 3, cols. 1–3 ($\times 10^3$) for the 1915, 1920, and 1925 periods, and 1970, 1975, and 1980 periods, respectively. Lines B.1 and B.2 use the dependent variable in Table 3, cols. 4–6. The entries on lines A.4 and B.4 are the changes in the predicted values over the relevant period from the regressions in cols. 3 and 6 of Table 3, respectively. These are found by multiplying the change in each economic variable by its coefficient from the indicated Table 3 regression and summing. The r = coefficient of correlation between actual and predicted change.

The answers are summarized in Table 5 (for the Senate) and Table 6 (House). These tables reveal a profound change in regional voting patterns over the course of the century, and they show that *nearly all* this change can be attributed to changed economic interests. The basic facts about voting patterns are on the first three lines of each panel of each table (positive values denote support for liberal policies). In the early part of the century, support for liberal economic policy came mainly from the South, and opposition from the Northeast and Pacific states. Today, these alignments are exactly reversed (see panel A). Party alignments changed similarly: the number of northern Democrats grew and victory for southern Republicans became conceivable. But a profound change in the same direction also occurred within each party. This is shown on the first three lines of panel B in the tables. In the early twentieth century, regional differences within parties were relatively small (line B.1).[22] But in the sixty years after World War I, the northern members of both parties grew more liberal and the southerners more conservative (B.3). These within party changes have accounted for a substantial part of the overall change in Congress (compare the standard deviations in col. 10 for lines A.3 and B.3) and have produced considerable regional differences within parties today.

[22] But economic differences were relatively large. The explanation for this strange pairing which is consistent with my previous results rests on the opposition of history and economics. Prior to World War I, the strong northern economic interest in conservative economic policy and southern interest in liberal policy clashed with opposite historical tendencies (see the pattern of the dummy variables in Tables 3 and 4, cols. 3 and 6). At the level of Congress as a whole (line A.1) the economic interest dominated, but within parties the two forces offset each other (B.1).

TABLE 6–CHANGE IN FREQUENCY OF LIBERAL VOTING IN THE HOUSE, c.1920–c.1975, REGIONS

Liberal Voting Measure	Region and Voting Measure $\times 10^3$									*SD* Across Regions
	NE (1)	MA (2)	ENC (3)	WNC (4)	SA (5)	ESC (6)	WSC (7)	MT (8)	PAC (9)	(10)
A. Within House (No Party Adjustment)										
1. Actual c.1920	−227	−108	−90	−13	171	230	231	−35	−144	169
2. Actual c.1975	183	108	−54	−85	−80	−91	−35	−117	73	105
3. Change (2.–1.)	410	216	36	−98	−252	−321	−266	−82	217	253
									(r = .98)	
4. Predicted Change	340	210	121	−84	−187	−195	−247	−90	143	207
5. Residual (3.–4.)	70	6	−85	−14	−65	−126	−19	8	74	67
B. Within Parties (Net of Party Effect)										
1. Actual c.1920	−65	−13	2	73	−23	25	−5	−28	20	39
2. Actual c.1975	170	120	45	−10	−183	−130	−180	−58	65	129
3. Change (2.–1.)	235	133	43	−83	−160	−155	−175	−30	45	142
									(r = .98)	
4. Predicted Change	182	115	66	−46	−103	−107	−136	−50	78	114
5. Residual (3.–4.)	48	18	−23	−37	−67	−48	−39	20	−33	36

Note: See Table 5. The same techniques used to generate data in that table are used here. Predicted and actual values come from regressions like those in Table 3 rather than Table 4, i.e., the House data are grouped into regions and regressions like those in Table 3 (using 9 regional dummies) are used to generate the coefficients of the economic variables which are then used to calculate the predicted changes on lines A.4 and B.4. The data are from 4-period averages with 1910, 1916, 1920, and 1926 comprising the first period, and 1966, 1970, 1976, and 1980 the second.

I think that the most noteworthy finding of this paper is the remarkably close degree to which these profound political changes can be attributed to changes in economic interest. This is seen by comparing lines 3 and 4 of each panel in both Tables 5 and 6. The "predicted changes" on line 4 are obtained from the coefficients of the economic variables in the regressions in col. 3 or col. 6 in Table 3 or 4, and the change in those variables from the early to late twentieth century. There are 36 predicted changes in Tables 5 and 6 (two voting measures for each House for each of nine regions). Only 1 of these 36 disagrees with the sign of the actual change. Only 2 deviate from the actual change by more than half the standard deviation of the actual change. None of the correlation coefficients between these actual and predicted changes is below .95.[23]

The results are similar if the longer period is divided into two subperiods of roughly equal length centered on the end of World War II. This is done in Table 7. These subperiods have somewhat different characteristics—more stable regional party alignments and slower erosion of regional economic differences in the pre-World War II period—but the simple economic model is able to rationalize most of the political change in both subperiods: of the 72 pairs of actual and predicted changes in Table 7, the signs agree in 63 cases. The correlation coefficient between these two variables never falls below .80 in the eight series in the table, and averages .86. The substantive message of Table 7 is that the South's move away from liberal policies and the North's move toward them is not compressed into the recent period when regional party alignments began chang-

[23]The regressions in Table 4 and associated data permit comparison of actual and predicted changes across the 48 states. These state-level data are also highly correlated (.82 for both the within-House and within-party measures). In spite of the greater "noise" in these state-level data, the positive correlation holds even after the very large regional elements are removed: the correlation of the actual with predicted *deviations* of state changes from the regional means is .37 for within-House data and .49 within-party (both are significant).

TABLE 7—ACTUAL AND PREDICTED CHANGES IN FREQUENCY OF LIBERAL VOTING IN CONGRESS, TWO SUBPERIODS (c.1920–1945, 1945–1975) REGIONS

Type of Change and Period	Region and Voting Measure × 10^3									Correlation of Actual and Predicted
	NE (1)	MA (2)	ENC (3)	WNC (4)	SA (5)	ESC (6)	WSC (7)	MT (8)	PAC (9)	(10)
A. Within House of Congress										
1. Senate, 1920–45										
a. Actual Change	194	136	−10	−80	−156	−61	−138	74	77	.92
b. Predicted Change	114	34	−12	−45	−103	−42	−49	−15	67	
2. Senate, 1945–75										
a. Actual Change	185	71	233	145	−187	−266	−153	−164	116	.82
b. Predicted Change	175	155	121	−76	−11	−232	−180	−48	48	
3. House, 1915–45										
a. Actual Change	112	83	−6	−102	−63	−154	−91	128	252	.84
b. Predicted Change	131	46	−34	−38	−101	−45	−102	8	127	
4. House, 1945–75										
a. Actual Change	341	164	84	59	−274	−246	−268	−239	11	.93
b. Predicted Change	227	168	147	−49	−106	−168	−160	−81	23	
B. Within Parties										
1. Senate, 1920–45										
a. Actual Change	104	63	4	−80	−51	45	−23	31	31	.80
b. Predicted Change	76	15	−11	−30	−71	−27	−25	−4	38	
2. Senate, 1945–75										
a. Actual Change	92	179	85	55	−134	−196	−92	−82	62	.90
b. Predicted Change	107	107	77	−56	−77	−165	−124	−22	41	
3. House, 1915–45										
a. Actual Change	116	63	54	−6	−120	−143	−93	83	97	.82
b. Predicted Change	73	26	−18	−21	−55	−24	−56	3	70	
4. House, 1945–75										
a. Actual Change	120	67	3	−57	−73	−23	−110	−107	−37	.86
b. Predicted Change	126	92	81	−27	−58	−92	−89	−46	12	

Note: Actual and predicted changes are computed in the same manner as in Tables 5 and 6 (see their Notes), but for two subperiods. Three-term averages of the relevant data are computed for an "early," "middle," and "current" period as follows. The early period is an average of 1910, 1916, and 1920 data for the House and 1915, 1920, and 1925 data for the Senate. The middle period is an average of 1940, 1946, and 1950 data for the House and 1940, 1945, 1950 for the Senate. The current period is 1970, 1976, and 1980 for the House and 1970, 1975, and 1980 for the Senate. The changes shown above as 1920–45 or 1915–45 are middle minus early data; the 1945–75 changes are current minus middle data.

ing. The economic forces underlying these shifts and the political response to them are palpably evident long before this and continue to work essentially up to the present.

So far I have forced on the data a model in which economic change is the only source of political change. An alternative story would be that the noneconomic regional preferences, which I have so far assumed to have remained unchanged, have in fact changed as well. That alternative cannot be ignored in light of the seemingly massive and long-lasting political realignments engendered by the New Deal. Could not, for example, the post-New Deal rise of labor unions and ethnic and racial constituencies in the North have been responsible for the shift toward liberal politics in that region? Table 7 provides part of the answer—the shift was going on before the New Deal. But I sought a more formal test. Instead of *assuming* that the regional effects in the columns 3 and 6 regressions of Tables 3 and 4 never changed, I added a set of post-New Deal regional dummies; each = 1 for an observation on a particular region for 1940 and after, 0 otherwise. The coefficients of these post-New Deal regional dummies show the extent to which regional voting patterns (net of the effects of the economic variables) *changed* from the

TABLE 8—MEASURES OF RELATIVE IMPACT OF ECONOMIC AND HISTORICAL CHANGE ON POLITICAL CHANGE IN CONGRESS, WORLD WAR I TO PRESENT

	Correlation Coefficient Actual Change vs. Components of Predicted Change		*Beta* Coefficients for Components of Predicted Change	
Type of Change	*ECON* (1)	*HIST* (2)	*ECON* (3)	*HIST* (4)
A. Within Houses				
1. Senate	.92	.43	.92	.22
2. House	.93	−.30	1.17	.39
B. Within Parties				
1. Senate	.95	.08	.99	.30
2. House	.93	.74	.71	.47

Note: Data are from the unrestricted model (see text) in which changes in both economic and historical forces are permitted to affect voting behavior; this model permits the coefficients of the regional dummies to change between the pre- and post-1940 periods. Accordingly, the predicted (*PRED*) change in voting behavior for any region has two components in this model: (a) *ECON* = change due solely to changes in economic variables (i.e., holding constant any shift in the coefficient of the regional dummy). (b) *HIST* = change in the coefficient of the regional dummy.

Cols. 1 and 2 show the correlation coefficient between actual change and each component of *PRED*. Cols. 5 and 6 show the contribution of each component to *PRED* in standard deviation (*SD*) units. For example line A.1 says "a region where *ECON* is 1 *SD* above the mean will have a *PRED* .92 *SD* above the mean," etc. These *beta* coefficients are calculated by dividing *SD* of *ECON* and of *HIST* by *SD* of *PRED*. Panel A data are for changes in behavior across all congressmen without regard to their party. Panel B refers to changes within parties.

pre- to the post-New Deal period. Test of the null hypothesis (that the set of regional coefficients changed) generates statistics with an F-distribution as follows:[24]

F(Senate)	$= 0.85$	$(d.f. = 9{,}105)$
F(Senate, within parties)	$= 2.09$	$(9{,}105)$
F(House)	$= 1.64$	$(9{,}114)$
F(House, within parties)	$= 2.30$	$(9{,}114)$

$F_{.05} \approx 2.0 \qquad F_{.01} \approx 2.6$

These numbers imply rejection of the null hypothesis at the 1 percent level for all four regressions, but acceptance at 5 percent for two of them. So, the evidence for ("noneconomic") shifts in regional political preferences is weak, and this provides justification for the restricted model in which economic change alone drives political change.

[24]For the House, I am testing the hypothesis that coefficients of *regional* dummies in regressions using *regional* data changed over time. My computer program could not perform a similar test on the coefficients of the 48 state dummies.

Shifts in regional preferences seem quantitatively, as well as statistically, insignificant. We already know that the restricted model (i.e., only economic change matters) explains virtually all of the change in political behavior. So, the only way that changes in regional preferences could plausibly be an important source of political change would be for the unrestricted model to "reapportion" explanatory power from the economic variables to those measuring the shift in regional preference. However, Table 8 shows that this is not what the unrestricted model does. The data are based on the two separate elements of the change in voting behavior predicted by the unrestricted model—the *ECON* element (the sum of coefficients of the economic variables times the changes in these variables over time) and the change in historical (*HIST*) preferences (the change in the coefficients of the regional dummies from the pre- to post-1940 period). The first two columns show that, standing alone, *ECON*

TABLE 9—CORRELATION COEFFICIENTS BETWEEN STANDARD DEVIATIONS OF LIBERAL VOTING MEASURES IN CONGRESS AND OF ECONOMIC VARIABLES; 1910 OR 1915 TO 1980; ACROSS REGIONS OR STATES

Standard Deviation of Liberal Voting Measure	Correlation Coefficient with *SD* of		
	MFG (1)	*PCI* (2)	*URB* (3)
A. Unadjusted (for Persistent Regional or State Differences)			
1. Within			
a. Senate	.10	−.14	.03
b. House: Regions	.48	.45	.53
: States	.01	−.01	.01
2. Within Parties in			
a. Senate	−.86	−.85	−.80
b. House: Regions	−.67	−.74	−.67
: States	−.74	−.66	−.73
B. Adjusted for Persistent Regional or State Differences			
1. Within			
a. Senate	.82	.76	.75
b. House: Regions	.87	.85	.84
: States	.83	.87	.82
2. Within Parties			
a. Senate	.77	.86	.76
b. House: Regions	.82	.83	.85
: States	.71	.76	.79

Note: Each entry is a simple correlation coefficient between a time-series of standard deviations of an economic variable across regions or states and a time-series of a *SD* of one of the political liberalism measures analyzed previously (see Tables 3 and 4 Notes). Each House time-series has 15 observations and each Senate time-series has 14 observations. See text for method of calculating voting measures used in panel B.

is the much more *reliable* guide to the data than *HIST*; indeed the correlations in column 1 are not much lower than those previously reported for the restricted model. The last two columns show that *ECON* is quantitatively the much more important element in the change in behavior predicted by the unrestricted model.

All of this implies that what I had before merely labeled the "persistent historical" element in political behavior (i.e., the coefficients of the regional dummies in Table 3 and 4) really is persistent. This focuses more sharply my earlier challenge to future research: an economic explanation of these regional differences will have to uncover economic differences among regions that *have not changed much since the beginning of the century*. The regional homogenization evident in so many dimensions of economic activity make this a formidable challenge.

VI. A Reexamination of Previous Puzzles

I showed (Section II) that political differences among regions in Congress have not declined along with economic differences. That seeming anomaly is restated in panel A of Table 9. This shows simple correlations between the *dispersions* of the economic and the various congressional voting measures I have been analyzing. If there is a simple connection between narrowing economic and political differences among states, these should be consistently positive, and they obviously are not. However, if there are persistent regional elements in political behavior, there should be a positive correlation over time between the dispersions of economic variables and of voting measures *net of the persistent regional element*. Accordingly, I subtracted the appropriate coefficient of the regional (or state) dummy in Table 3

TABLE 10—REGRESSIONS OF SENATE VOTING PATTERNS, ON ECONOMIC AND POLITICAL CHARACTERISTICS STATES, 1979

	Coefficient of								
Dependent Variable	*HH INC* (1)	*MFG* (2)	*URB* (3)	*MET* (4)	*SPEND/TAX* (5)	*DEMS* (6)	*HIST* (7)	R^2	*SE*
1. *NTUA*									
a					.13	20.8	39.8	.41	10.1
					1.8	4.6	3.4		
(Table 1)					(−.01)			(.24)	(11.4)
c	−5.96	−.46	−.59	.27		16.0	92.3	.59	8.7
	3.7	2.1	2.2	2.0		3.9	5.4		
(Table 1)	(.48)	(.03)	(−.25)	(.18)				(.28)	(11.4)
2. *ADA*									
a					.08	32.3	95.0	.68	12.4
					.8	5.8	6.6		
(Table 1)					(−.28)			(.34)	(17.5)
c	−3.85	.08	−.52	.02		28.6	148.7	.80	10.2
	2.1	.3	1.7	.1		6.1	7.5		
(Table 1)	(.65)	(.88)	(.03)	(−.13)				(.51)	(15.7)

Note: The regressions follow the same format as counterparts in Table 1, panel B, except for the addition of *HIST* as an independent variable. *HIST* is a vector of the coefficients of the state dummy variables in the regression in col. 6, Table 4. That is, *HIST* measures the historical liberalism of a state's representatives (I have no state-specific data for senators) relative to their party mean over the whole 1910–80 period after accounting for economic variables. For comparison, the regression coefficients and summary statistics from the corresponding Table 1 regression are shown in parentheses on the lines labeled (Table 1). See Table 1 for definitions and sources of all variables other than *HIST*.

and 4 from each measure of voting behavior and recomputed the simple correlations between the dispersions of the adjusted political and economic variables. These are in panel B of Table 9 and they are all strongly positive.

The persistent regional element has to be removed to reveal this tandem decline of political and economic differences because the conflict between history and economics described in Section IV has abated overtime. Early in the century, the South's historical conservatism and the North's liberalism (see the regional coefficients in cols. 3 and 6 of Tables 3 and 4) clashed sharply with the redistributive interests entailed by southern poverty and northern affluence. This clash made regional differences in political behavior smaller than otherwise. However, the clash and its restraining influence on regional political differences, has diminished with the relative economic rise of the South.[25] Interestingly, my data imply that regional political differences will *grow* in the future even as the economic element of these differences diminishes.[26]

The weak or even perverse relationship between voting and economic interest often found in contemporary Congresses (see the discussion surrounding Table 1) can also be clarified by my results. They suggest that, for example, wealthy areas sometimes produce

[25]To see formally why the variance of political behavior (S^2) across regions need not decline along with the variance of the economic (+any random) element (E^2) of that behavior, note that $S^2 = E^2 + H^2 + 2rEH$, where H^2 = variance of the historical (i.e., time-invariant) element across regions, and r = correlation between the historical and economic (+random) elements across regions at time t. Then

$$\frac{dS}{dt} = \frac{1}{S}\left[\frac{dE}{dt}(E + rH) + \frac{dr}{dt}\cdot EH\right].$$

For most periods in the data, $dE/dt < 0$, but $r < 0$, so the sign of dS/dt is indeterminate.

[26]In the notation of fn. 25, E^2 is now so small that, with $r < 0$, $(E^2 + 2rEH) < 0$ and $S^2 < H^2$. Therefore, if E continues to approach zero, over time S^2 will rise toward H^2. For the Senate, in the period 1970–80, $S = .128$ while $H = .185$. The latter figure is about equal to the maximum s observed in this century.

liberal congressmen because the pull of history can overcome the push of interest. This is more likely today, when differences in interest are smaller, than it has been in the past. If this explanation is correct, then some adjustment for history should bring the role of interest in contemporary voting into sharper focus. This adjustment is made in Table 10 which adds history, in the form of a vector of the coefficients of state dummy variables from the column 6, Table 4 regression, to some of the regressions in Table 1.[27] The coefficient of this *HIST* variable (col. 7) is uniformly positive and significant, which is further testimony to the durability of these sectional differences. More important, with the addition of *HIST*, the coefficients of the economic variables change in the "right" direction from their Table 1 values (shown in parentheses); that is, income, manufacturing and urbanization, which are negatively correlated with the benefits of redistribution, have algebraically smaller and usually negative coefficients in Table 10.[28] Similarly, the partial correlation of liberal voting with the benefits from redistribution (lines 1a and 2a) is positive in Table 10, while it was strangely negative in Table 1.

VII. Summary and Conclusions

The evidence in this paper is consistent with a model in which congressional agents act as if they are maximizing a utility function like

$$U = F(L, W), \tag{5}$$

where L = the number or frequency of liberal votes cast,[29] and W = their principals' wealth per capita.

The agent's choice of L affects W via the political redistribution process, and the usual first-order conditions are

$$F_L/F_W = -dW/dL, \tag{6}$$

the "price" of a liberal vote in terms of W. That price depends on the nature of redistribution—it will be positive in some constituencies and negative in others. Accordingly, (6) has an interior solution only where liberalism is a costly good (F_L, $-dW/dL$; both >0) or a productive bad. My data imply that where liberalism seems to be a good (the North) it has historically been costly, and where it is a bad (the South) it has been productive. This amounts to invoking a "tastes" category to permit F_L to be nonzero in the same way that, say, an analyst of the market for rock music might have to invoke tastes to "explain" why some pay to hear it and others pay to avoid it. But to invoke tastes is also to challenge future research. In this case, the challenge is to uncover objective forces which can reduce the importance of my particular tastes category.

Meanwhile, I have followed the traditional path in utility analyses of choice, that of focusing on the effects of changes in constraints. The characteristics of American political redistribution suggest that, in general, the shadow price of a liberal vote ($-dW/dL$) rises with relative wealth, and, when that price rises, we expect fewer such votes to be "bought." That expectation is strongly confirmed by my data, as is the underlying assumption that the tastes in question are stable. I have shown that this conventional economic model is powerful enough to explain substantially all of the major political realignments among regions in this century. The economic convergence of congressional constituencies has gradually lowered the price of a liberal vote to north-

[27] These coefficients describe historical preferences in the *House* and the regressions in Table 10 describe voting in the *Senate*. This is done because I do not have comparable state-level data for the Senate. I also ran the regressions in Table 10 with the relevant regional *HIST* of the Senate. These were qualitatively similar to those in Table 10, but none explained the data as well as its Table 10 counterpart.

[28] The coefficient of *MET*, however, moves in the "wrong" direction, in that the crude evidence in Table 1 implies that residents of SMSAs lose from redistribution.

[29] At this level of generality, one has to be agnostic about whether L generates utility for the agents or principals or both.

ern congressmen and lowered the price of a conservative vote to southerners. This elemental fact is sufficient to explain 1) why the once conservative North has become liberal and why the opposite occurred in the South; 2) why once more-or-less homogeneous parties have become regionally divided, with northern members of either party now more liberal and southerners now more conservative than their party average, and 3) why Democrats have gained "market share" in the North and lost it in the South. Since the process of economic convergence appears not to have run its course, the strong suggestion of my results is that these political trends will continue. The seemingly paradoxical result predicted by my data is that Congress will become more sharply divided regionally as their constituencies converge economically.

REFERENCES

Buchanan, James and Tullock, Gordon, *The Calculus of Consent*, Ann Arbor: University of Michigan Press, 1962.

Downs, Anthony, *An Economic Theory of Democracy*, New York: Harper & Row, 1957.

Fenno, Richard, *Home Style: House Members in their Districts*, Boston: Little Brown, 1978.

Fiorina, Morris, *Representatives, Roll Calls & Constituencies*, Lexington: Lexington Books, 1974.

Kalt, Joseph, *The Economics and Politics of Oil Price Regulation*, Cambridge: Harvard University Press, 1981.

______ **and Zupan, Mark,** "Capture and Ideology in the Economic Theory of Politics," *American Economic Review*, June 1984, *74*, 279–300.

Kau, James and Rubin, Paul, "Self-Interest Ideology and Logrolling in Congressional Voting," *Journal of Law and Economics*, October 1979, *22*, 365–84.

______ **and** ______, *Congressmen, Constituents and Contributors*, Boston: Nijhoff, 1982.

Lee, Everett S. et al., *Population Redistribution and Economic Growth, United States, 1870–1950*, Vol. I., Philadelphia: American Philosophical Society, 1957.

Leven, Maurice, *Income in the Various States*, New York: National Bureau of Economic Research, 1925.

Mitchell, Edward J., "The Basis of Congressional Energy Policy," *Texas Law Review*, March 1979, *57*, 591–613.

Peltzman, Sam, "Constitutent Interest and Congressional Voting," *Journal of Law and Economics*, April 1984, *27*, 181–210.

Reynolds, Morgan and Smolensky, Eugene, *Public Expenditures, Taxes and the Distribution of Income*, New York: Academic Press, 1977.

Stigler, George, "The Theory of Economic Regulation," *Bell Journal of Economics*, Spring 1971, *2*, 3–21.

U.S. Department of Commerce, Bureau of the Census, *Historical Statistics of the United States*, Washington: USGPO, 1975.

______, *Statistical Abstract of the United States*, Washington: USGPO, various years.

U.S. Department of Labor, Bureau of Labor Statistics, *Handbook of Labor Statistics*, Washington: USGPO, 1967.

U.S. Internal Revenue Service, *Annual Report of the Commissioner*, Washington: USGPO, various years.

Short-Run Equilibrium Dynamics of Unemployment, Vacancies, and Real Wages

By CHRISTOPHER A. PISSARIDES*

In this paper, I study the dynamics of adjustment in a labor market, following an exogenous shock to the real value of output. Some of the stylized facts of business cycles, with which the predictions of the model are consistent, include first, real wages do not fully reflect fluctuations in the real value of labor's marginal product, so real profits fluctuate more than real wages. Second, unemployment responds to output shocks, but its response is slow. Finally, in countries where there are good data on vacancies, like Britain, we observe that vacancies respond more quickly to shocks and with greater amplitude than unemployment.

Several authors have constructed models to explain why output shocks are absorbed partly by real wages and partly by unemployment (the empirical regularity in the United States is discussed by Robert Hall, 1980). Implicit contract models (Costas Azariadis, 1979; Oliver Hart, 1983) have successfully explained why real wages may not reflect output shocks, and the models with asymmetric information and severance pay have also had some success in explaining fluctuations in unemployment. Bargaining models (Ian McDonald and Robert Solow, 1981) and efficiency wage models (Janet Yellen, 1984) appear to be more successful in explaining fluctuations in unemployment, but formalizations are still in their infancy. The models have not yet been subjected to the same scrutiny as implicit contract and earlier models.

One feature shared by all these models is that they are static. They explain how real wages and employment respond to shocks in a comparative-static framework but say nothing about the adjustment path from one equilibrium to the next. Also, the models say nothing about job vacancies, either in equilibrium or during the adjustment process. By contrast, this paper takes the view that by modeling job vacancies explicitly, one can learn more about the behavior of unemployment and real wages, both in equilibrium and during the adjustment to equilibrium. Thus, the model developed below is explicitly dynamic, and in it job vacancies play a critical role in the transmission of output shocks to real wages and unemployment.

A job vacancy indicates a willingness by a firm to hire a worker.[1] It is equivalent to unemployment of capital, so just as workers move between the states of employment and unemployment, jobs move between the states of occupancy and vacancy. I model the interaction of vacancies and unemployment by using ideas from equilibrium search theory, where there is continuous wage recontracting and perfect anticipation of the adjustment paths of all endogenous variables. Job vacancies enter the model via their influence on job contacts, which depend on the number of firms looking for workers. Some firms may not wish to hire and so they may not be actively engaged in the search process. Only firms with job vacancies are actively engaged in search, so the number of job contacts and

*London School of Economics, Houghton Street, London WC2A 2AE. Work on this paper was started at NBER's Summer Institute in July 1983 and later completed at the Industrial Relations Section of Princeton University; I thank both institutions for their hospitality and financial support. I also thank the referee and seminar participants at the universities of Boston, Chicago, Iowa, Wisconsin, and Yale for helpful comments.

[1]For more discussion of the concept of vacancies, with empirical evidence for the United States and Britain, see, respectively, Katharine Abraham (1983) and my paper with Richard Jackman and Richard Layard (1983). Earlier contributions incorporating vacancy-unemployment interactions (but without explicit micro models and with a peripheral role for wages) include Charles Holt and Martin David (1966) and Bent Hansen (1970).

the flow of workers out of unemployment depend on the number of job vacancies.

The static properties of the model closely resemble those of other models cited: real wages do not fully reflect output shocks and unemployment absorbs some of the shocks. In order to derive an adjustment path from this setup, I make use of the observation that vacancies respond more quickly to shocks than unemployment. When desired employment rises, firms open up new vacancies to indicate their willingness to hire more labor, and unemployment falls when labor takes up these vacancies. Similarly when desired employment falls, firms withdraw their vacancies, making it more difficult for unemployed workers to find jobs. Hence, changes in vacancies lead changes in unemployment. The critical assumption that I make in this paper is that vacancies are a nonpredetermined, fully flexible variable, whereas unemployment is predetermined, except for additions to it that result from upward movements in reservation wages. Wages are also nonpredetermined with continuous recontracting.

The assumption concerning vacancies is obviously extreme, since opening up new vacancies often requires the acquisition of new capital, which may not be readily available. However, if the existence of a job vacancy is a prerequisite to a hiring, vacancies must respond at least as quickly to shocks as unemployment, whatever the length of time required to open up a vacancy. The assumption that I make here enables the development of a simple model whose dynamic behavior is governed by this differential speed of response, and not by any time lags in the acquisition of capital.

I consider the effects of a multiplicative shock to the value of output and to capital costs, under the assumption of perfect foresight. Changes in unemployment result from successful job contacts and from job separations, making unemployment a relatively sluggish predetermined variable. Vacancies and wages adjust continuously so as to ensure that during adjustment the economy stays on a unique perfect-foresight adjustment path.

The assumptions of the model imply that the unique adjustment path is characterized by a constant vacancy-unemployment ratio and constant real wages. Hence these variables change only in response to news about shocks and not during the adjustment following the news. Moreover, real wages do not change by as much as the value of output and capital costs, inducing firms to change the number of job vacancies. Following this initial response to the shock, the number of job contacts changes, leading to simultaneous changes in vacancies and unemployment until the economy reaches its new equilibrium.[2] The model predicts that vacancies overshoot their long-run equilibrium value when expansionary or contractionary shocks are first observed. Following the initial impact of the shock, they change in the same direction as unemployment, tracing anticlockwise loops in vacancy-unemployment space.

The model also predicts that the time-series of unemployment, vacancies, and real wages will be characterized by asymmetries. The response of unemployment to a negative shock is faster than the response to a positive shock. The reason for this is that, when the shock is negative, reservation wages and profits do not fall by as much as actual wages and profits, leading to a number of immediate job separations. There are no corresponding immediate job matchings when the shock is positive, because the process that brings together firms and workers is time consuming. For the same reason, vacancies overshoot by less when the shock is negative, and mean observed wages respond less at the onset of a negative shock than at the onset of a positive one. The job separations when the shock is negative re-

[2] Thus, in time-series data, the model predicts that changes in productivity-corrected real wages should be small and uncorrelated, whereas unemployment should follow an autoregressive process. These predictions are consistent with the U.S. and U.K. data of Joseph Altonji and Orley Ashenfelter (1980), and with the U.S. real-wage data of Ashenfelter and David Card (1982).

move from the market low-wage jobs, raising the mean wage.[3]

Of crucial importance in the derivation of these results is the fact that labor has some alternative use to production, whose real return is not sensitive to expansionary or contractionary output shocks. This alternative use determines partly the workers' threat point in the wage bargain, introducing an element of inflexibility in wages.[4] Firms' threat points are determined partly by the cost of idle capital, which I assume varies in proportion to the value of output. This assumption, however, is not important for the qualitative impact of output shocks.

Sections I, II, and III develop the equations describing the behavior of unemployment, vacancies, and wages. Section IV derives the reservation prices and discusses job rejection. Section V brings the elements together and derives the unique perfect foresight path for this economy. Finally, Section VI describes the economy's response to a multiplicative output shock.

I. Unemployment

Consider an economy consisting of a fixed labor force (which is used as the normalizing constant) and of a variable number of jobs. At any point in time, a fraction $1-u$ of the labor force is employed and the remaining fraction u is unemployed and looking for a job. The variable u, as well as the other endogenous variables, are functions of time, but the time notation will be suppressed for convenience. Steady-state equilibrium values will be distinguished by a bar over the relevant variable. The number of jobs as a fraction of the labor force is $1-u+v$, where v denotes the job vacancies which are waiting for workers to arrive. It is assumed that no firm without a vacancy can take on a worker; that is, having a job vacancy is a prerequisite for participating in the search process that brings together jobs and workers.

There are frictions in the labor market, which make it impossible for all the unemployed to find jobs instantaneously. I shall not model explicitly the source of these frictions.[5] Instead, it will be assumed that the implication of these frictions is that the typical unemployed worker is faced with a probability p of making a contact with a firm with a job vacancy, the typical vacancy is contacted by a worker with probability q and both p and q lie strictly between 0 and 1. It is plausible to assume that these contact probabilities depend on the relative number of firms and workers engaged in job search. Writing θ for the ratio of vacancies to unemployment (v/u ratio) we assume that both p and q are continuous differentiable functions of θ, with $\partial p/\partial\theta \geq 0$ and $\partial q/\partial\theta \leq 0$.[6]

There is a simple relationship between p and q, because, by definition, the number of unemployed workers who make contacts with jobs must be equal to the number of vacancies that are contacted by workers. Thus, with a typical contact probability of p and with u job searchers, the number of workers who make contacts is up; similarly, the number of job vacancies that are contacted by workers is vq. Hence, $up = vq$, or, using the θ notation,

$$p(\theta) = \theta q(\theta).$$

The assumption that $\partial p/\partial\theta \geq 0$ and $\partial q/\partial\theta \leq 0$ implies that the elasticities of p and q

[3]Recently, Salih Neftci (1984) provided evidence for the asymmetric behavior of unemployment which is consistent with the predictions derived here.

[4]McDonald and Solow make a similar argument for wage inflexibility in their union bargaining model.

[5]The literature on imperfect information and job search discusses extensively the sources of these frictions. See, for example, Edmund Phelps (1972).

[6]Assuming that p and q depend on the ratio θ and not on the absolute levels of u and v imposes some homogeneity restrictions on the process of search which, in some models, have strong implications for the efficiency of equilibrium outcomes. See, for example, my 1984 paper, where I used a steady-state version of this model to study the efficiency of job rejection in the presence of search externalities. In this model, homogeneity simplifies the exposition without materially affecting the analysis. There are also some reasons to believe that homogeneity is a plausible restriction a priori; see, for example, Stephen Nickell (1979), where the probability p is successfully estimated as a function of the v/u ratio, and Hall (1979) where a homogeneous search process is explicitly derived.

with respect to θ must be less than 1 in absolute value.

Not all job-worker pairs are equally productive. I assume that when a worker and a vacancy come together they immediately establish their net output from a potential match. Let this be y units and suppose that y is a random drawing from a cumulative density function $F(y)$. The density function is identical for all jobs and workers, so the reason for different productivities is the difference in the efficiency of the job match, not in general skills or technologies. I assume that y is a constant flow per unit time and that it is parametric.

When a firm and a worker meet, they agree whether to form a job match, or whether to reject it and try again. All potential job matches with nonnegative surplus are made. The surplus is divided according to a wage function whose derivation I shall describe in Section III. For the moment, let x denote the productivity that yields zero surplus, so all jobs matches with productivity $y \geq x$ are made, whereas job matches with productivity $y < x$ are rejected. The reservation productivity x is a nonpredetermined choice variable.

The job matches that are made at each point in time equal the fraction of contacts that yield productivity at least as high as x. Let

$$a = \int_x^\infty dF(y).$$

Then, in a large market the number of job matches made is simply aup; the number of contacts made is up, and a is the fraction of acceptable contacts.

Additions to unemployment take place exogenously, at the flow rate $s(1-u)$. The jobs broken up are selected randomly, so each job-worker pair is faced with an exogenous separation probability s, regardless of the productivity of the match or the time it was formed. These assumptions, although strong, are needed to make the analysis tractable. Out of the steady state, separations may also take place endogenously as a result of changes in the reservation productivity x. The exogenous separation process assumed may be justified by appeal to firm-specific shocks that cancel out in aggregate, such as structural shifts in final demand or random obsolescence (and breakdown) of machines.

Unemployment changes in response to the flows in and out of jobs. The rate at which it changes is given by

$$\dot{u} = (1-u)s - a(x)p(\theta)u, \tag{1}$$

where a dot denotes a time derivative. The variables u, θ, and x are functions of time, but s is assumed to be constant. Equation (1) makes unemployment a sluggish variable, and it is one of the fundamental dynamic equations of the model.[7]

II. Vacancies

The number of job vacancies is determined by firms, in response to the expected profit from a new vacancy. For simplicity, I shall use the terms firm and job interchangeably; that is, it will be assumed that there are constant returns to scale without substitution possibilities between labor and capital after the installation of capital, so each firm may be modeled as having only one job. Intuitively, a job may be thought of as a machine that could be operated by one worker. Firms could acquire machines for a fixed rental, and in order to engage in the search process that leads to job matches, they must have an idle machine. The number of machines in existence depends on the expected profit from an extra machine, given the expected duration of a job vacancy and the profit from production. Machines can be brought into use, rented and scrapped instantaneously, so the number of job vacancies is a perfectly flexible nonpredetermined variable.

Let us assume that apart from the foregone profit that a firm suffers when it has a vacancy, it has to bear also a cost k per unit time. The cost k is assumed to be a flow in order to simplify the exposition; it represents

[7]Some exogenous changes may induce once-and-for-all step changes in unemployment, and at these points equation (1) is not defined. See Section V below.

the fixed cost of machines that has to be borne regardless of whether jobs are filled or not, and any other labor-recruitment costs that the firm may have. Since net output y is parametric, it may be thought, in general, as being net of the part of k that has to be borne when the machine is occupied. Other costs netted out of y (but not out of k) include the cost of raw materials.

The flow of profit from a job with productivity y is given by $y-w(y)$, where $w(y)$ is the wage rate. Wages are chosen by firms and workers after they meet according to a Nash-bargaining rule, in a way described in the next section. Let V be the asset value of a vacancy (i.e., an idle machine) and J^e the expected asset value of a filled job, and suppose there is a perfect capital market with fixed interest rate r. Then, the asset value of a vacancy must yield a net return that is equal to the yield from the vacancy, plus the expected capital gain from finding a worker:

$$(2) \qquad rV-\dot{V}=-k+aq(J^e-V).$$

In capital market equilibrium, the free-market yield rV must be equal to what the firm expects to get from an idle machine: an appreciation $\dot{V}$, a net cost k, and a probability aq of filling the job, and so of making a capital gain of J^e-V. The expected capital value of a filled job is the conditional expectation

$$J^e=E(J(y)|y\geq x),$$

where $J(y)$ is the expected asset value of a job with productivity y. In general, the asset value of a job will depend on the productivity of the job match, and since only job matches at least as productive as x are accepted, the best a firm can do when calculating the expected profit from a job match is to take the conditional expectation of $J(y)$.

Firms will acquire machines for as long as $V>0$, and will scrap them when $V<0$. Hence, if machines can be bought and sold without lags, equilibrium implies $V=0$. This yields an equilibrium restriction on J^e:

$$(3) \qquad J^e=k/a(x)q(\theta).$$

Equation (3) holds at all times, both in and out of steady-state equilibrium.

The asset value of an occupied machine that produces output y, $J(y)$, satisfies a condition similar to (2). Recalling that the profit from the occupied machine is $y-w$ and that the probability of losing a worker is s, we obtain

$$(4) \qquad rJ(y)-\dot{J}(y) = y-w(y)+s(V-J(y)).$$

Since $V=0$, this simplifies to

$$(5) \qquad \dot{J}(y)=-(y-w(y))+(r+s)J(y).$$

Conditions (3) and (5) may be solved for θ in terms of a and $w(y)$, which are also endogenous. With knowledge of θ, we can obtain u and v from (1) and the definition of $\theta=v/u$. However, it is more convenient to treat θ as the unknown rather than v, and I shall be doing this in the development of the model. The next task is to specify the wage function $w(y)$ and the reservation productivity x that determines a. As I will show, they may be expressed as functions of θ and the exogenous variables, so their determining conditions are, like (3) and (5), independent of the levels of u and v.

III. Wages

Wages are assumed to be determined by Nash bargains between the meeting firm and worker and to be perfectly flexible; that is, there can be continuous renegotiation and recontracting.[8] The (generalized) Nash rule says that the surplus from a job match is divided between the firm and the worker according to a fixed parameter β, $0\leq\beta\leq 1$. The surplus enjoyed by firms when the productivity of the job is y is $J(y)-V$, which in equilibrium is simply equal to $J(y)$. In order to calculate the surplus enjoyed by workers, we need to derive expressions for the worker's

[8]Similar rules for wages, but in a steady-state equilibrium only and without variations in productivities, were also discussed by Peter Diamond (1982).

net worth (asset value) when occupied and when idle.

Let U be the worker's asset value when he is unemployed and $W(y)$ be his asset value when he is employed in a job producing output y. Then, if b denotes the worker's return when he is not producing (including any unemployment benefits and net of out-of-pocket search costs), U satisfies

$$(6) \qquad rU-\dot{U}=b+ap(W^e-U).$$

where b is assumed to be parametric,[9] and ap is the transition probability for workers, as already defined. The conditional expectation $W^e=E(W(y)|y\geq x)$, and it gives the worker's expected net worth from a job, given that all jobs with productivity below x are rejected.

If a worker is in a job with productivity y, his net worth $W(y)$ satisfies

$$(7) \quad rW(y)-\dot{W}(y)=w(y)+s(U-W(y))$$

This equation has the same interpretation as (6): the worker receives $w(y)$ from the job (which will in general depend on y) and faces a probability s of returning to unemployment with reward U.

The worker's net surplus from a job match with productivity y is $W(y)-U$. I assume that the worker gets a fraction β of the total surplus $W(y)+J(y)-U-V$. Hence, wages are fixed so as to satisfy the condition

$$(8) \quad W(y)-U=(\beta/(1-\beta))(J(y)-V).$$

In equation (8), U acts as the worker's threat point in the wage bargain, and V acts as the firm's threat point. The best that each side can do if they fail to agree on a sharing rule is to search optimally for another match. Then, the parameter β may be interpreted as a coefficient measuring bargaining strength independently of the relative position of the two sides' threat points. In the symmetric Nash case, examined by Peter Diamond, $\beta=1/2$.

To derive an equation for wages subtract (6) from (7), and rearrange to obtain

$$(9) \quad (r+s)(W(y)-U)-(\dot{W}(y)-\dot{U}) \\ =w(y)-b-ap(W^e-U).$$

Substitute now (8) into (9), noting that (8) implies

$$\dot{W}(y)-\dot{U}=(\beta/(1-\beta))(\dot{J}(y)-\dot{V}).$$

The result is

$$(10) \quad (\beta/(1-\beta))[(r+s)(J(y)-V) \\ -(\dot{J}(y)-\dot{V})] \\ =w(y)-b-ap(W^e-U).$$

But (2) and (4) imply

$$(r+s)(J(y)-V)-(\dot{J}(y)-\dot{V}) \\ =y+k-w(y)-aq(J^e-V).$$

Hence (10) becomes, after rearranging,

$$(11) \quad w=(1-\beta)b+\beta(y+k) \\ +(1-\beta)ap(W^e-U)-\beta aq(J^e-V).$$

Thus, workers receive a payment b, plus a fraction β of the net surplus from the job $y+k-b$, plus an amount depending on perceptions of the gains from employment elsewhere, minus an amount depending on perceptions of the firm's gains from recruiting another worker to the job. If perceptions are correct then (8) implies $W^e-U=(\beta/(1-\beta))(J^e-V)$, and in equilibrium $V=0$, J^e is given by (3), and by definition $q=p/\theta$. Hence the wage equation (11) becomes

$$(12) \quad w(y)=(1-\beta)b+\beta(y+\theta k).$$

Equation (12) is the generalized Nash wage equation, holding out of steady-state equilibrium, but only when perceptions of wages

[9] That is, we ignore variations in the intensity of search, whereby a worker may increase his contact probability p by spending more time searching (thus lowering b).

and profits elsewhere are correct and when the expected profits from a vacancy are zero. In the steady state, this equation becomes

$$\bar{w}(y) = (1-\beta)b + \beta(y + \bar{\theta}k).$$

Interestingly, $w(y)$ does not depend on the rate of growth of u or of any other variable. This property simplifies the short-run dynamic analysis of the model and has some strong implications for its behavior. Also, with correct perceptions and the zero profit condition $V = 0$, what goes on in the rest of the market (for example, the distribution of productivities or the rate of structural change) influences $w(y)$ only indirectly, through the vacancy-unemployment ratio θ. If $V \neq 0$, the mean productivity y^e and other parameters also influence $w(y)$ directly through a third term, which vanishes only if $u = v$ (see my earlier paper).

With $\dot{V} = 0$, any exogenous change that raises the vacancy-unemployment ratio raises wages because it improves the worker's threat point in the wage bargain, relative to the firm's. For similar reasons, an increase in the worker's return from nonmarket activities b, or an increase in the firm's vacancy costs k, also increase wages for given θ. The former improves the worker's threat point, whereas the latter deteriorates the firm's threat point. But when account is taken of the effect of b and k on θ (see equation (20) below), then the effect of k on wages is reversed: an increase in the cost of idle capital reduces wages, so as to compensate firms for the higher nonlabor costs. Finally, a higher match-specific productivity y, given the distribution of productivities $F(y)$, implies higher wages and profits, with wages rising at a rate β and profits at a rate $1-\beta$. If the higher productivity is not specific to the match but general, in equilibrium workers receive more than a fraction β of the increase, because a general increase in productivities increases the vacancy-unemployment ratio (see Section VI).

IV. Job Rejection

Now, I am in a position to describe the choice of reservation productivity by firms and workers, and so close the model.

Suppose a firm and a worker meet and discover that the productivity of their match is equal to y. If the firm accepts to form the match, it will enjoy an expected return $J(y)$, whereas if it rejects it, it will search again and so enjoy the expected returns from a vacancy, V. Hence firms will be willing to accept all matches which satisfy $J(y) \geq V$, or, since in equilibrium $V = 0$, $J(y) \geq 0$. It follows that the reservation productivity x satisfies

$$J(x) = 0. \tag{13}$$

Similarly, if a worker accepts a job match with productivity y, he will enjoy returns $W(y)$, whereas if he rejects it he will go back to net worth U. Hence workers accept all matches that satisfy $W(y) \geq U$, giving a reservation x satisfying $W(x) = U$. By the equilibrium wage condition (8), the x that solves (13) also satisfies the worker's condition: hence firms and workers agree about the reservation job, and to obtain its properties we need only consider (13).

The equilibrium x is immediately obtained from (5) by substituting in it $y = x$ and making use of (13). Then $w(x) = x$, and using the wage equation (12), we get

$$x = b + (\beta/(1-\beta))\theta k. \tag{14}$$

Equation (14) closes the system. It is in unfamiliar form, compared with reservation-wage formulas derived elsewhere, because it already incorporates the equilibrium condition $V = 0$. Firms and workers, of course, ignore the effect that their actions have on equilibrium, so it is not true to say that when they choose their reservation productivity they simply look at b, β, θ, and k, and calculate it according to (14). But after they make their choice of reservation productivity, the creation and closure of jobs in the market as a whole will ensure that their chosen x behaves according to (14). The form of (14) is the most convenient one for our purposes because, like the wage rate in (12), x is expressed as a linear function of θ, and of no other endogenous variable.

Equation (14) suggests that any exogenous change that increases the vacancy-unemployment ratio θ increases the reservation pro-

ductivity. Thus, despite the fact that both workers and firms search and they both agree on which job matches to reject, higher job availability is associated with more job rejection. The asymmetry that gives rise to this result is the assumption that the size of the labor force is fixed, whereas the number of jobs is variable. This puts firms at a disadvantage in the job-matching process, so in equilibrium the marginal job with productivity x pays the entire product as wages. In partial equilibrium models of search where only workers reject jobs, higher job availability induces workers to select higher reservation wages, so the same relation holds between job availability and job rejection.[10]

V. Equilibrium and Short-Run Dynamics

Equations (3), (5), (12), and (14) contain four unknowns: $J(y)$, θ, $w(y)$, and x. We may eliminate the wage equation (12) by substituting into (5), to obtain

$$(15)\quad \dot{J}(y)=-(1-\beta)(y-b)+\beta\theta k+(r+s)J(y).$$

Hence, equations (3), (14), and (15) may now be solved for the three unknowns $J(y)$, θ, and x.

I first express (15) as a linearized differential equation in θ, by making use of (3) and (14). Using bars to denote steady-state values, I write (3) as

$$(16)\quad J^e=\frac{k}{a(\bar{x})q(\bar{\theta})}-\frac{kq'(\bar{\theta})}{a(\bar{x})q(\bar{\theta})^2}(\theta-\bar{\theta})-\frac{ka'(\bar{x})}{a(\bar{x})^2q(\bar{\theta})}(x-\bar{x}).$$

But from (14), $x-\bar{x}=(\beta/(1-\beta))k(\theta-\bar{\theta})$, so substituting into (16), we get

$$(17)\quad J^e=\frac{k}{a(\bar{x})q(\bar{\theta})}-\frac{k}{a(\bar{x})q(\bar{\theta})}\times\left[\frac{q'(\bar{\theta})}{q(\bar{\theta})}+\frac{a'(\bar{x})}{a(\bar{x})}\frac{\beta}{1-\beta}k\right](\theta-\bar{\theta}).$$

Taking now conditional expectations of (15), we obtain

$$(18)\quad \dot{J}^e=-(1-\beta)(y^e-b)+\beta\theta k+(r+s)J^e.$$

Substution of J^e and $\dot{J}^e$ from (17) into (18) yields the equation in θ:

$$(19)\quad \dot{\theta}=\left[(1-\beta)(y^e-b)-(r+s)\frac{k}{aq}\right]\times\left[\frac{q'}{q}+\frac{a'}{a}\frac{\beta}{1-\beta}k\right]^{-1}\frac{aq}{k}-(r+s)\bar{\theta}+\left[r+s-\beta aq\left(\frac{q'}{q}+\frac{a'}{a}\frac{\beta}{1-\beta}k\right)^{-1}\right]\theta.$$

The arguments of a and q have been omitted for notational convenience, it being understood that all coefficients are evaluated at $\bar{\theta}$.

Equation (19) is a fixed-coefficients differential equation in θ. Since $q'(\theta)<0$, $a'(x)<0$, the coefficient of θ in (19) is positive, making it an unstable equation. Hence, the only perfect foresight solution for θ is its steady-state value $\bar{\theta}$. If $\theta\neq\bar{\theta}$, the system will diverge on an explosive path. Thus, if any parameter of the system changes, θ will adjust immediately to its new steady-state value. This steady-state value is given by (19) when $\dot{\theta}=0$ and $\theta=\bar{\theta}$:

$$(20)\quad \frac{r+s}{a(\bar{x})q(\bar{\theta})}+\beta\bar{\theta}=(1-\beta)\frac{y^e-b}{k}.$$

Equation (20) gives the value of θ both in and out of steady-state equilibrium. By the wage equation (12) and the reservation productivity equation (14), both wages and the reservation productivity are also always at

[10] If $V=0$ is not built into the derivation of the reservation productivity level, the formula giving x is very similar to the standard reservation-wage formula derived in partial models, except that it depends on both the worker's and the firm's costs and returns, and not only on the worker's. See my earlier paper. Then, higher θ values are not necessarily associated with more job rejection, because with a fixed number of jobs ($V\neq 0$) firms and workers are treated with full symmetry.

their steady-state values, $\overline{w}(y)$ and $\bar{x}$. But unemployment is a predetermined variable and adjusts according to (1):

$$\dot{u}=s-[s+a(\bar{x})p(\bar{\theta})]u. \tag{21}$$

During the economy's adjustment, both x and θ are at their steady-state values so, since θ is defined as the v/u ratio, vacancies change along with unemployment, in such a way as to ensure that θ is constant. This simply requires that vacancies change in the same proportion as unemployment; that is, $\dot{v}/v=\dot{u}/u$, or, using the θ notation, $\dot{v}=\bar{\theta}\dot{u}$.

Now, it was argued in Section I that additions to unemployment may take place also endogenously, following a rise in the reservation productivity x for given distribution $F(y)$. This possibility introduces a further element into the unemployment dynamics of equation (21), which, because of the behavior of x, is easy to deal with. Thus, if following a parametric change, some jobs are no longer viable, the firm and worker are involved separately, creating an immediate inflow into unemployment. Following this impact change, unemployment changes smoothly according to (21).

In contrast to increases in unemployment, there are no impact changes when the reservation productivity falls. It is not possible for unmatched firms and workers to come together and form jobs which they had rejected earlier, other than through the search process. Thus, unemployment behaves asymmetrically on impact: following a rise in x relative to the distribution of y, there is an immediate rise in unemployment followed by smooth adjustment; following a fall in x relative to $F(y)$, unemployment changes smoothly from the start.

I shall illustrate the properties of this model by considering its response to a shock that may reasonably be argued to correspond to what we typically observe over a business cycle.

VI. Response to Output Shocks

Consider a multiplicative real shock to the output from each job, y, and to the cost of a vacancy k. Thus, output and capital costs rise in equal proportion, but the worker's alternative return b does not rise. This causes a differential change in the net returns that workers and firms get from employment, producing some real responses of the endogenous variables.

It could be argued that this change in relative costs and returns is the main channel through which the business cycle affects the labor market: there is a general rise in output prices and costs, unaccompanied (in the short run at least) by changes in the value of the workers' alternative return (compare McDonald and Solow, p. 896). As I show, the response of real wages, number of jobs and unemployment to this kind of shock simulates the observed stylized changes of these variables over a typical cycle. In particular, real wages fluctuate by less than output and other costs, and, as a consequence, the v/u ratio rises in the peak and falls in the trough, and unemployment falls in the peak and rises in the trough.[11]

Suppose then y and k depend on a parameter h, and write

$$y(h)=(1+h)y; \quad k(h)=(1+h)k.$$

I evaluate the effect of (unanticipated permanent) changes in h at the point $h=0$.[12] It is

[11] I show below that the system is neutral to a multiplicative shift in y, k, and b, so the response to a multiplicative shift in y and k is the same (but of opposite sign) as the response to a multiplicative shift in b. This seems to suggest that the effects of changes in unemployment benefits, which can be represented by changes in b, are similar to the effects of the cycle. Two caveats should be noted here. First, cyclical shocks are short-lived, whereas the effects of changes in benefits are thought to be permanent. This does not make a qualitative difference to modeling the two kinds of shocks, but it does make a difference when it comes to an interpretation of the data. Second, if a multiplicative shock to y and k persists (like, for example, technological changes do) it is unlikely that b will remain unaffected. The value of workers' time in nonmarket activities now is not what it was in the last century, surely because of large technological improvements in production and rises in the standard of living.

[12] The modeling of cyclical shocks as a series of unanticipated permanent shocks in a perfect foresight model, such as the one of this paper, may be criticized on the grounds that agents will eventually realize the regularity of the shocks and incorporate them into their decisions. The motivation for the modeling of shocks in

convenient to derive first the effect of changes in h on y^e and a. To do this, we need to write expressions for $a(h)$ and $y^e(h)$ for a new distribution function $g(y)$, which lies to the right of the old distribution $f(y)$ and it is also more spread out. Consider first $a(h)$.[13] By definition,

$$a(h)=\int_{x(h)} g(y)\,dy,$$

whereas at $h=0$, $a=\int_x f(y)\,dy.$

To compare the two expressions, let us express $g(y)$ in terms of $f(y)$, by noting that the change from $f(y)$ to $g(y)$ is achieved by multiplying y by $(1+h)$. The fact that densities must integrate to 1 implies

$$1=\int g(y)\,dy=c\int f\left(\frac{y}{1+h}\right)dy,$$

where c is a normalizing constant. Let $z=y/(1+h)$, hence

$$1=c\int f(z)(1+h)\,dz=c(1+h),$$

and so $c=1/(1+h)$.

Now, returning to the expression for $a(h)$, we have

$$a(h)=\frac{1}{1+h}\int_{x(h)} f\left(\frac{y}{1+h}\right)dy,$$

or, changing the variable of integration to z,

$$a(h)=\int_{x(h)/(1+h)} f(z)\,dz.$$

By a similar argument, we can derive

$$y^e(h)=a(h)^{-1}\int_{x(h)/(1+h)} z(1+h)f(z)\,dz.$$

Hence, differentiating $a(h)$ and $y^e(h)$ with respect to h, and evaluating at $h=0$, we obtain

$$\text{(22)}\qquad \frac{\partial a}{\partial h}=-\left(\frac{\partial x}{\partial h}-x\right)f(x);$$

$$\text{(23)}\qquad \frac{\partial y^e}{\partial h}=y^e-\frac{1}{a}(y^e-x)\frac{\partial a}{\partial h}.$$

The acceptance probability changes only if the reservation productivity does not change in the same proportion as the exogenous productivities. Similarly, the mean conditional productivity changes by the same proportion as each individual productivity only if the acceptance probability is unaffected by the change.

To see now the crucial role played by wages in inducing real responses to changes in h, consider the unemployment equation (21). Unemployment will respond to changes in h if $a(x)$ or $p(\theta)$ respond to them. But in the derivation of x in (14) we saw that in equilibrium $x=w(x)$. Therefore, if wages change in the same proportion as productivities, x will also change in the same proportion, and from (22) we get that $a(x)$ will be unaffected by the change.

Similarly, we derived θ from (3) and (5), by noting that along the perfect foresight path $\dot{\theta}=0$. Hence $\dot{J}=0$, and (3) and (5) give

$$q(\theta)=\frac{(r+s)k}{a(x)(y^e-w^e)}.$$

If wages change in the same proportion as y and k, $a(x)$ is unaffected, so from (23) and an equivalent expression for w^e, we obtain that y^e and w^e also change in the same proportion. Hence, with $a(x)$ constant, and y^e, w^e, and k changing by the same proportion, $q(\theta)$ and θ must be unaffected by the change in h.

It follows then that neither unemployment nor vacancies will change if wages change in the same proportion as output and costs. But

this paper is, first, that forecasts of the timing and intensity of shocks are not good, even though agents may know that some shocks will occur, so shocks contain most of their news value when they occur. Second, the paths of unemployment, vacancies, and real wages between turning points are dominated by endogenous adjustments following substantial shocks that disturb the equilibrium, rather than by responses to frequent serially correlated shocks.

[13] I am indebted to a referee for suggesting the method of derivation of $a(h)$ that follows.

the wage equation (12) implies that wages will change in the same proportion as y and k only if b changes in the same proportion too, for then θ is constant. Hence, an equiproportional change in b, y, and k is absorbed entirely by wages, and has no employment effects. But if the worker's alternative return b is not as sensitive to the shocks as productivity and capital costs, wages will not respond fully to the shocks and there will be real effects.

To derive the full effects of the shock when b is fixed, I return to the condition determining θ. This is given by equation (20), rewritten as (noting that $q = p/\theta$)

$$(24)\quad \frac{(r+s)\theta}{a(h)p(\theta)} + \beta\theta = (1-\beta)\frac{y^e(h)-b}{k(1+h)}.$$

Differentiation with respect to h yields

$$(25)\quad \left[\frac{r+s}{ap}(1-\eta)+\beta\right]\frac{\partial\theta}{\partial h} = \frac{(1-\beta)b}{k},$$

where η is the elasticity of $p(\theta)$ and it is less than 1. Hence $\partial\theta/\partial h > 0$ unambiguously.

The reservation productivity rises too, since, from (14),

$$(26)\quad \frac{\partial x}{\partial h} = \frac{\beta}{1-\beta}\left(k\frac{\partial\theta}{\partial h} + \theta k\right) > 0.$$

But the increase in x is not as big as the increase in y and so the acceptance probability rises. Substitution of (26) into (22) yields

$$\frac{\partial a}{\partial h} = \left(-\frac{\beta}{1-\beta}k\frac{\partial\theta}{\partial h} + b\right)f(x),$$

and by making use of (25), we obtain

$$(27)\quad [(r+s)(1-\eta)/(ap)+\beta]\,\partial a/\partial h = (r+s)(1-\eta)bf(x)/(ap) > 0.$$

The results that I have derived so far take place instantaneously, as soon as the change in h is realized. The equations that I have used hold both in and out of steady-state equilibrium, so the instantaneously observed changes in θ, x, and a are also the final changes. By contrast, unemployment does not change on impact if a increases: it rises if a decreases, as the now unacceptable jobs break up. But following these impact changes, unemployment changes in response to the changes in a and $p(\theta)$ according to (21). If $h > 0$, unemployment starts falling towards a new steady-state equilibrium, since both contact probabilities $p(\theta)$ and acceptance probabilities $a(x)$ rise. If $h < 0$ unemployment rises.

Returning now to the wage equation (12), we find that the response of wages to the shift parameter h is

$$(28)\quad \frac{\partial w(y)}{\partial h} = \beta(y+\theta k) + \beta k\frac{\partial\theta}{\partial h}.$$

The change in wages is proportionally less than the change in y or k, despite the term $\beta k\,\partial\theta/\partial h$. By making use of (25), it can be shown that

$$\frac{\partial w(y)}{\partial h}\frac{1}{w(y)} < \frac{\partial y}{\partial h}\frac{1}{y} = 1.$$

Profits, as a result, rise by more than in proportion, so firms increase the number of job offers. Also, because the rewards available to firms and workers who reject jobs are relatively less than before, they are more willing to accept jobs than before. Thus, in terms of wages and output, what we observe in the market is an increase in the value of output accompanied by a partial response of real wages, leading to more job vacancies and more job acceptances.

Following these changes unemployment starts falling, as both job contacts and job acceptances increase. The lower unemployment leads to a decrease in the number of jobs taken up (since the number of searchers falls with every fall in unemployment) and an increase in the number of job separations (because the employment rate increases). Eventually a new equilibrium is reached with lower unemployment, higher wages and profits, and higher v/u ratio.

Vacancies increase at first on impact so as to drive θ up to its new equilibrium level. Then, as unemployment starts to fall vacancies fall too, since, with a constant θ, $\dot{v} = \bar{\theta}\dot{u}$.

Thus, in the first phase of expansion, vacancies overshoot their new equilibrium value. It is not possible to say whether the final equilibrium value of vacancies is above or below the old equilibrium value. At the point $\dot{u}=0$, we have

$$\partial\bar{v}/\partial h=\bar{u}\partial\theta/\partial h+\theta\partial\bar{u}/\partial h, \tag{29}$$

where, from (21),

$$\frac{\partial\bar{u}}{\partial h}=-\frac{p\bar{u}}{s+ap}\frac{\partial a}{\partial h}-\eta\frac{pa\bar{u}}{\theta(s+ap)}\frac{\partial\theta}{\partial h}. \tag{30}$$

Hence, substituting into (29), we obtain

$$\frac{\partial\bar{v}}{\partial h}=-\frac{p\bar{u}\theta}{s+ap}\frac{\partial a}{\partial h}+\frac{s\bar{u}+(1-\eta)pa\bar{u}}{s+ap}\frac{\partial\theta}{\partial h}. \tag{31}$$

This cannot in general be signed, if only because $\partial a/\partial h$ depends on $f(x)$, which makes a comparison between the two terms in (31) inconclusive. Equation (31) points to the fact that there are two effects on equilibrium vacancies. A positive effect from the increase in θ, and a negative one from the increase in the acceptable job matches. In general we cannot say which effect dominates, but if $f(x)$ is small, the positive effect through θ will dominate.[14]

The behavior of unemployment and vacancies in response to a positive and negative shock h is shown diagrammatically in Figure 1. The $\dot{u}=0$ locus slopes downwards under the assumption

$$-ap'(\theta)+p(\beta/(1-\beta))kf(x)<0.$$

That is, when the effect of vacancies on unemployment through job availability dominates the effect through the increase in

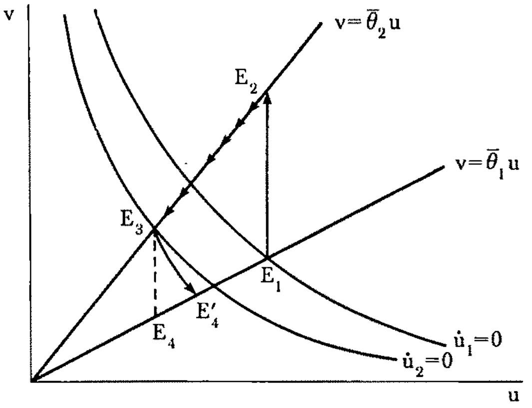


FIGURE 1. THE EFFECT OF MULTIPLICATIVE OUTPUT SHOCKS ON VACANCIES AND UNEMPLOYMENT

reservation productivities (holding output and all other variables constant; see the discussion in the preceding paragraph and in fn. 14). Under the reasonable restriction $p''(\theta)\le 0$, the locus is also convex to the origin if the job availability effect dominates, but this is not important for our purposes. The stable trajectory (saddlepath) is the $\theta=\bar{\theta}$ locus, shown by the straight line through the origin.

Suppose the economy is initially at E_1, and a positive multiplicative shock takes place. Then, on impact, θ rises (say from θ_1 to θ_2) pivoting the stable trajectory to the left. The economy jumps on impact from E_1 to E_2, on the new trajectory. The $\dot{u}=0$ locus also shifts inwards, because of the increase in θ and $a(x)$, say from $\dot{u}_1=0$ to $\dot{u}_2=0$. At point E_2 both vacancies and unemployment start falling towards the new long-run equilibrium E_3, where unemployment is lower, but vacancies may be lower or higher than at E_1.

Now if, starting at E_3, there is a negative shock of a similar magnitude the curves shift back, so the new equilibrium is at E_1. But now the point reached on impact is not E_4, but some point to the right of it, like E_4'. The reason for this is that at the time of the negative shock, the number of acceptable job matches falls. So there are some jobs which are no longer acceptable, and these break up immediately, leading to an immediate increase in unemployment. Following this ini-

[14] This kind of ambiguity frequently arises in models of search, except that it is discussed only in relation to unemployment. Here it does not arise on the unemployment side because the cause of a rise in θ is a positive output shock, which increases the acceptance rate despite the increase in θ. The common conjecture is that the effect through job availability normally dominates.

tial impact, the economy travels smoothly from E_4' to E_1, along the stable trajectory. Thus during contractions the economy's adjustment is faster than during expansions, and the cycle that it traces about the $\dot{u}=0$ curve is of less amplitude.[15]

Finally, it should be noted that observed mean wages and output in the market change less than the wage or output of a typical job, y. The observed mean output is y^e and the observed mean wage is

$$(32) \qquad w^e = (1-\beta)b + \beta(y^e + \theta k).$$

It follows immediately from (23) that the proportional change in y^e is less than 1: the number of acceptable jobs increases (if $h > 0$) adding some low productivity jobs to those already in the market. As a result, there is another dampening effect on mean wages, besides $(1-\beta)b$. Like the effect of a on unemployment, this dampening effect operates with a lag in an expansion, leading to a mild overshooting of mean productivity and wages in the first stages of expansion. But in a contraction it operates without lag, so the response of mean wages and productivity on impact is less in contraction than in expansion.

VII. Conclusions

Early search theory was effectively criticized for its exclusive reliance on supply-of-labor responses and for not providing an adequate theory of wages (see, for example, James Tobin, 1972). The model presented in this paper meets both of these criticisms by treating supply and demand symmetrically, and by relying on vacancies as a link between output shocks and unemployment. Thus, wages are endogenously determined by bargaining at the individual level and they are fully flexible, so firms and workers always agree about which jobs to accept and which to reject. But the supply of vacancies is determined exclusively by profit-maximizing firms, so workers may find themselves unemployed for lengthy periods of time because of limited vacancy availability.

The comparative static predictions of the model are similar to those of contract theories, and other theories that developed partly in response to the criticisms of early search theory. A shock to labor's marginal and average product is absorbed partly by real wages and partly by employment. The role of some alternative return to labor that is insulated from the shock is crucial in this prediction. But unlike other recent theory, the model can shed light on the adjustment paths of the endogenous variables. Search considerations suggest that unemployment is a sluggish (predetermined) variable and if we add to this the assumption of nonpredetermined vacancies (and also of nonpredetermined real wages and reservation prices) we can derive unique perfect-foresight adjustment paths for all the endogenous variables. The role of vacancies in pushing the economy to its unique perfect-foresight path is crucial: at the onset of expansions and recessions, vacancies have to overshoot their equilibrium value and return subsequently to equilibrium, tied to unemployment. By contrast, real wages and reservation prices change only in response to news, and even then they do not fully reflect the output shocks. Thus, in contrast to unemployment which is highly serially correlated, changes in (productivity-corrected) real wages should be small and uncorrelated, unless the shocks are serially correlated. Moreover, the model predicts that the response of unemployment to a negative shock should be faster than its response to a positive shock. Both of these predictions are consistent with the time series data on real wages and unemployment (Altonji and Ashenfelter; Neftci).

A possible objection to the model is that although its predictions conform to the commonly observed anti-clockwise loops in vacancy-unemployment space in Britain, it predicts too much response to news. Thus, empirically vacancies do not seem to overshoot their equilibrium value to the extent

[15] The $\dot{u}=0$ curve is often referred to as the Beveridge curve. The argument of this paper suggests that the Beveridge curve is not stable over the cycle, as is normally assumed. The evidence concerning the British curve in the 1960's and 1970's is discussed in my paper with Jackman and Layard. Hansen also derived the anti-clockwise loops in $u-v$ space, but he relied on *ad hoc* adjustment equations with a speculative component in vacancy decisions, giving him the lead of vacancies over unemployment.

that the model predicts, and some may argue that real wages sometimes behave like a predetermined variable. However, it would need a much more careful study of the data than now available to establish the validity of these propositions; for example, there is no analysis (to my knowledge) of the effects of news in labor markets, of the kind that one finds in recent research on asset markets. But in addition, more sluggishness may be incorporated into the model by realizing that there may be predetermined elements in some aspects of vacancy decisions. For example, the extent to which firms can open up new vacancies in response to news may be limited by their ability to acquire new capital. Then, in expansions the initial jump in vacancies may be checked, and changes in the vacancy-unemployment ratio may continue after the revelation of news. Real wages would then change sluggishly during adjustment, despite continuous recontracting, because of their dependence on the vacancy-unemployment ratio.

A formal model incorporating these elements would have to deal with the asymmetries that are likely to arise in booms and recessions. The availability of capital will not introduce sluggishness into vacancy decisions in a recession, though it might do in a boom. This paper has made a start at the formal analysis of short-run dynamics by making the simple assumption of predetermined unemployment vs. nonpredetermined vacancies. Generalizing this assumption may shed more light on the extent of fluctuations in response to news in labor markets, but it is not likely to alter the general patterns of adjustment derived.[16]

[16] If output shocks are to some extent anticipated, as they might if they are associated with regular cycles, the response of vacancies and real wages to them will also be less pronounced and more spread out. Less than perfect anticipations have the effect of spreading out the revelation of news, with predictable results.

REFERENCES

Abraham, Katharine, G., "Structural/Frictional vs. Deficient Demand Unemployment: Some New Evidence," *American Economic Review*, September 1983, *83*, 708–24.

Altonji, Joseph and Ashenfelter, Orley, "Wage Movements and the Labour Market Equilibrium Hypothesis," *Economica*, August 1980, *47*, 217–45.

Ashenfelter, Orley and Card, David, "Time Series Representations of Economic Variables and Alternative Models of the Labour Market," *Review of Economic Studies*, Special Issue 1982, *49*, 761–82.

Azariadis, Costas, "Implicit Contracts and Related Topics: A Survey," in Z. Hornstein et al., eds. *The Economics of the Labour Market*, London: HMSO, 1979.

Diamond, Peter A., "Wage Determination and Efficiency in Search Equilibrium," *Review of Economic Studies*, April 1982, *49*, 217–27.

Hall, Robert E., "A Theory of the Natural Unemployment Rate and the Duration of Employment," *Journal of Monetary Economics*, April 1979, *5*, 153–69.

______, "Employment Fluctuations and Wage Rigidity," *Brookings Papers on Economic Activity*, 1:1980, 91–124.

Hansen, Bent, "Excess Demand, Unemployment, Vacancies and Wages," *Quarterly Journal of Economics*, February 1970, *84*, 1–23.

Hart, Oliver, "Optimal Labour Contracts under Asymmetric Information: An Introduction," *Review of Economic Studies*, January 1983, *50*, 3–35.

Holt, Charles C. and David, Martin, H. "The Concept of Job Vacancies in a Dynamic Theory of the Labor Market," in *The Measurement and Interpretation of Job Vacancies*, NBER Other Conference Series, No. 5, New York: Columbia University Press, 1966.

Jackman, Richard A., Layard, Richard and Pissarides, Christopher A., "On Vacancies," Centre for Labour Economics Discussion Paper No. 165, London School of Economics, 1983.

McDonald, Ian M. and Solow, Robert M., "Wage Bargaining and Employment," *American Economic Review*, December 1981, *71*, 896–908.

Neftci, Salih N., "Are Economic Time Series Asymmetric over the Business Cycle?," *Journal of Political Economy*, April 1984, *92*, 307–28.

Nickell, Stephen J., "Estimating the Probability of Leaving Unemployment," *Econometrica*, September 1979, *47*, 1249–66.

Phelps, Edmund S., *Inflation Policy and Unemployment Theory: The Cost Benefit Approach to Monetary Planning*, New York: W. W. Norton, 1972.

Pissarides, Christopher A., "Efficient Job Rejection," *Economic Journal Conference Papers*, 1984, *94*, 97–108.

Tobin, James, "Inflation and Unemployment," *American Economic Review*, March 1972, *52*, 1–18.

Yellen, Janet L., "Efficiency Wage Models of Unemployment," *American Economic Review Proceedings*, May 1984, *74*, 200–05.

International Trade and Labor Migration

By WILFRED J. ETHIER*

Footloose labor is prominent in the contemporary international economy. Migrant workers have accounted for up to a quarter of the Swiss labor force, and 10 per cent of that of northern Europe generally. A number of American industries now depend heavily upon migrant labor (both legal and illegal) from Mexico and the Caribbean. South Asia supplies a large part of the labor for oil-financed Middle Eastern economic development. Pakistan, for example, now earns almost three-quarters as much foreign exchange from the remittances of temporary migrants abroad as from conventional exports; China has begun formally to contract the export of labor services.[1]

For years, a number of European countries have administered formal guest-worker systems, and it has become commonplace to argue that such arrangements have helped insulate native populations from the rigors of industrial fluctuations and have also allowed the preservation of import-competing industries that would have succumbed completely to foreign competition without a supply of cheap migrant labor. But note the possibility of a latent contradiction between these two alleged benefits: preserving jobs for native workers in the face of an economic downturn by sending migrants home requires the two types of labor to be *substitutes*, but preserving domestic jobs by hiring cheap migrant workers requires *complementarity*. It is evident, then, that formal analysis is necessary to understand these host-country issues.

Although migrant labor has generated a large descriptive literature, formal economic treatment is scanty, especially in comparison with that accorded international capital movements. Two basic approaches may be noted. The first is concerned with the brain drain. The characteristic feature here is the coincidence of labor movements with that of human capital. Since the present paper is by contrast motivated by the large-scale movements of temporary migrants with few skills, I shall not allude further to this literature.

The second approach has been simply to introduce international labor mobility into standard factor-endowment models of international trade.[2] This results, for the most part, in treating migration symmetrically with international capital mobility. There can be no doubt that this approach is valuable. Useful insights have emerged, and the standard trade models have done all that can reasonably be asked of them. But, at the same time, certain central characteristics of contemporary experience are fundamentally alien to this approach.

At the risk of some oversimplification, one might isolate the following stylized facts as characteristic[3] of a preponderant share of modern migration.

*Professor of Economics, University of Pennsylvania, 3718 Locust Walk/CR, Philadelphia, PA 19104. This paper has benefited from seminar discussions at the universities of Pennsylvania and Rochester, the Institute for International Economic Studies in Stockholm, the Stockholm School of Economics, the Bijeenkomst Vakgroepen International Economies at Erasmus University on May 21, 1982, the 1982 NBER Summer Institute, and Columbia, Tel Aviv, and Yale universities. Helpful comments were also contributed by Eitan Berglas, Rob Feenstra, Gene Grossman, Lars Svensson, Chuck Wilson, and several anonymous referees.

[1]An excellent general treatment of labor migration may be found in Michael Piore (1979). See U. Hiemenz and K. W. Schatz (1979) for details about foreign workers in Germany, Ariadne Marshall (1973) for the Netherlands, C. Kennedy-Brenner (1979) for France, A. F. Corwin (1978) for Mexican migrants in the United States, J. S. Birks and C. A. Sinclair (1980) for the Middle East, and K. C. Zachariah and J. Conde (1981) for West Africa.

[2]See Bertil Ohlin (1967, ch. XVI), W. Max Corden (1955), Harry Johnson (1965, ch. III, part III), Murray Kemp (1969, ch. 9), James Markusen and James Melvin (1979), Avinash Dixit and Victor Norman (1980, pp. 146–49), Jagdish Bhagwati and Richard Brecher (1980), Brecher and Bhagwati (1981), and Bhagwati (1979).

[3]See the books mentioned in fn. 1, plus further references cited in those works.

(*i*) *Migration is regarded as temporary*. Migrants typically intend to work abroad only long enough to accumulate nest eggs to establish themselves back home, and host countries typically issue temporary work permits. Although both sometimes discover that "you can't go home again," the common view of migration as temporary is a distinctive characteristic of its role in the host-country economy. The employment arrangements of native labor are by contrast generally regarded as much more permanent.

(*ii*) *Migrant labor and native labor are imperfect substitutes in production*. It is often said that migrants hold jobs which native workers would not accept, and this seems to be crucial to the political acceptance of immigration in host countries. But, as noted above, complete complementarity would preclude attempts to use migrants to shield native workers from the business cycle.

(*iii*) *Migrants accept unpredictable variations in employment conditions more readily than do native workers*. The latter typically possess a degree of job security not accorded migrants.

(*iv*) *Migrant labor seems to be in practically unlimited supply at exogenously determined* (*but sometimes changing*) *terms*. Once they have established connections with source-country labor pools, host countries generally experience little difficulty in recruiting all the migrants they are willing to accept. When changing conditions close off one source of supply, other sources are quickly found.

(*v*) *Host countries typically experience much international capital mobility as well as labor mobility*. Both direct investment and portfolio investment transactions tend to be large. But general conclusions are not obvious about either the directions or net amounts of physical capital movements.

Although some of these characteristics can indeed be embodied into traditional trade models in quite straightforward fashion, others clearly cannot, and the stylized facts (*i*)–(*v*) on the whole describe a quite different world. Thus a new theoretical approach is needed. The present paper attempts to develop a theory that embodies the five characteristics and yet remains consistent with standard factor-endowments theory in the sense that the latter emerges as a special case. I then use the new theory to analyze the role of migration in host countries, addressing, for example, the questionable relation between the two frequently alleged benefits referred to above.

I. International Equilibrium

A. *Migrant Labor and Traded Goods*

A small country, call it Switzerland, produces a traded good, say chocolate. Chocolate is only a minor component of the Swiss consumption bundle: assume for simplicity that all chocolate is exported. These exports pay for imports of a composite traded good, which is consumed, and for the services of migrant labor.

Chocolate is produced by native labor and migrant labor, which are imperfect substitutes for each other. I assume that chocolate output is given by $Nf(m)$, where N denotes the employment of native labor, m the ratio of migrant to native labor employed, and f is a neoclassical production function (so that $f' > 0$ and $f'' < 0$).

Switzerland is too small to influence the relative prices of her imports, which can accordingly be treated as a single good, but the country is a sizable source of supply in the market for chocolate. Thus the price of chocolate (in terms of imports, i.e., Switzerland's commodity terms of trade) depends upon Swiss supply. It also depends upon the state of nature; for simplicity I assume there are only two possible states, each with probability one-half of occurrence. Thus equilibrium in each state is described by

$$X_A D(P_A) = eNf(m_A), \tag{1}$$

$$X_B D(P_B) = Nf(m_B). \tag{2}$$

A subscript refers to the state of nature, $D(P_i)$ denotes the state-invariant component of world demand for Swiss chocolate, X_i is the random variable, and P_i denotes the price of Swiss chocolate in terms of the numeraire: imports. N is the employment of native labor in state B, and e denotes the

ratio of state A employment of natives to N. I label as A the state resulting in lower native employment, so $e \leq 1$. Migrant labor is in perfectly elastic supply at an exogenous (but possibly state-dependent) wage, and firms can decide their employment of migrants after the state of nature is realized. Firms maximize expected profit and so in each state hire migrants to ensure

$$(3) \qquad f'(m_A) = v_A/P_A.$$

$$(4) \qquad f'(m_B) = v_B/P_B,$$

where v_i denotes the exogenous migrant wage in state i.

Equations (2) and (4) jointly determine the state B terms of trade P_B and production technique m_B, provided that N is given exogenously. Equations (1) and (3) yield a relation between P_A, e, and m_A. Solving out m_A yields

$$(P_A/e)(d_e/dP_A) = -\rho_A/(1-\theta_{MA}).$$

In this expression $\rho_A = \theta_{MA}\sigma_A + (1-\theta_{MA})\eta_A$, where σ_i, η_i, and θ_{Mi}, respectively, denote state i elasticity of substitution between native and migrant labor $(-f'(m_i)[1-(m_i f'(m_i)/f(m_i))]/m_i f''(m_i))$, elasticity of demand for Swiss chocolate $(-P_i D'(P_i)/D(P_i))$, and migrant labor's distributive share $(m_i f'(m_i)/f(m_i))$. Next, denote by a_i the value of the marginal product of native labor in chocolate production in state i:

$$a_i = P_i[f(m_i) - m_i f'(m_i)].$$

Note that a_B will be determined by (2) and (4), given N, but that a_A depends upon the actual combination of e and P_A. The latter cannot be specified independently of the native labor market, about which I have as yet said nothing.

Relations (1), (3), and the definition of a_A imply

$$(5) \qquad \hat{e} = -\rho_A \hat{a}_A,$$

where a circumflex denotes proportional change. Since e denotes the ratio of state A employment of natives in the chocolate industry to state B employment, it can be interpreted as a measure of employment volatility: lower values of e indicate more native employment volatility. Similarly, with a_B set by equations (2) and (4), lower values of a_A indicate more volatility across states of the marginal productivity of native workers in the chocolate industry (and thus, in a sense, of the degree of Switzerland's comparative advantage in chocolate production). Equation (5) then says that the international markets in which Switzerland deals present that country with a basic tradeoff: by accepting more volatility in the degree of comparative advantage, chocolate employment can be rendered less volatile. This tradeoff is depicted by the International Markets schedule, labeled *IM* in Figure 1. This curve, which from (5) has an elasticity of ρ_A, shows for each value of a_A that level of e consistent with equilibrium in the two international markets (chocolate and migrant labor) in which Swiss firms deal.

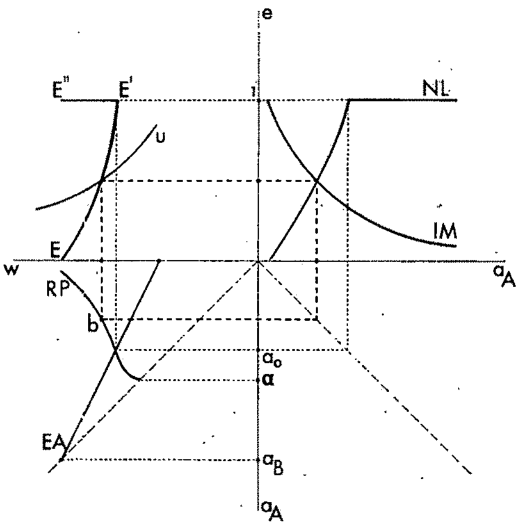


FIGURE 1

B. *The Substitution of Migrants for Native Workers*

The terms of the tradeoff along the International Markets schedule are measured by the basic substitution elasticity ρ_A: the greater the degree of substitutability, the greater the

reduction in employment volatility that can be "purchased" by an increase in comparative advantage volatility. Note that ρ_A is made larger by larger values of both the elasticity of production and the elasticity of demand for chocolate, and that a rise in the relative importance of migrants in production, as measured by θ_{MA}, will raise or lower ρ_A as the elasticity of substitution is greater than or less than the elasticity of demand.

These elasticities are also crucial in determining how the *location* of the International Markets schedule relates to the stochastic process confronting Switzerland. This process is summarized by the values of (v_A, X_A) and (v_B, X_B). Consider how changes in v_A and X_A shift the *IM* schedule, with (v_B, X_B) held fixed as a point of reference.

Suppose, first, that X_A rises. Equations (1) and (3) reveal that if $\hat{e} = \hat{X}_A$ (a circumflex denotes proportional change), this will be consistent with unchanged P_A, m_A, and, therefore, a_A. Thus the International Markets schedule shifts up in the proportion $\hat{X}_A$. Given the degree of comparative advantage in state A, a rise in world chocolate demand allows an equiproportional rise in native employment, and in migrant employment as well. The ability to hire migrants allows the productivity of native workers to be maintained, free of the penalty of diminishing returns, regardless of the degree of substitutability between the two types of labor.

Suppose, next, that v_A increases instead. Equation (3) reveals that, if m_A remains unchanged, $\hat{P}_A = \hat{v}_A$. Thus from the definition of a_A, $\hat{a}_A = \hat{P}_A = \hat{v}_A$ as well. Also, equation (1) requires that $\hat{e} = -\eta_A \hat{P}_A = -\eta_A \hat{v}_A$. Thus the effect of an increase $\hat{v}_A$ in state A migrant wages can be depicted as a rightward shift of the *IM* schedule in the amount $\hat{v}_A$ accompanied by a downward shift of $\eta_A \hat{v}_A$. Now we can see from (5) that, for a movement *along* the *IM* schedule, a rightward movement of $\hat{v}_A$ must be accompanied by a downward movement of $\rho_A \hat{v}_A$. Thus the net effect of the rise in v_A must be to shift up the International Markets schedule if and only if $\rho_A \hat{v}_A > \eta_A \hat{v}_A$, which is equivalent to $\sigma_A > \eta_A$. This reflects that fact that a rise in the state A cost of migrant labor influences native employment in two ways. On the one hand, natives are substituted for the now-more-expensive migrants; high values of σ_A obviously facilitate this substitution. But the rise in v_A also raises the cost of chocolate production, and thus the price that Swiss firms must charge. This will reduce world demand for Swiss chocolate and thereby tend to depress native employment; high values of η_A obviously magnify this effect. Thus it is not surprising that, given a_A, higher migrant wages v_A correspond to higher native employment levels e if and only if $\sigma_A > \eta_A$.

The present model in a very simple way gives the degree of substitutability between native labor and migrants very definite roles to play: setting the terms of the tradeoff between employment volatility and terms of trade volatility, and helping to position the International Markets schedule. To see where the economy actually operates on this tradeoff requires consideration of the native labor market, to which I now turn.

C. *The Native Labor Market*

This paper distinguishes migrants from native workers in two crucial ways. The previous subsection focused on the assumption that the two types of labor are imperfect substitutes in production. Now I focus on the second difference: the way each type relates to chocolate firms. Migrants can be hired freely by domestic firms at prevailing wages once the state of the world becomes known, but firms and native workers have a more long-term relationship. In particular, they establish a (possibly implicit) contractual relationship before the state of nature becomes known.[4]

In light of stylized fact (v) noted in the introductory remarks, I assume that Switzerland is integrated into the world capital market in the sense that chocolate firms are owned by capitalists with globally diversified portfolios. Further, I suppose that this capital market affords complete opportunities to

[4] This subsection draws upon the early implicit contract literature. See for example, Costas Azariadis (1975).

hedge Swiss risks, so that the integration is manifested by risk neutrality on the part of chocolate firms.

The contractural arrangements between firms and native workers potentially can take alternative forms. Two cases seem worthy of mention. In what I term the *European mode*, the contract fully specifies the transaction that will take place irrespective of the state of nature that actually materializes: a firm that takes on a worker commits itself to employ him at a predetermined wage, come hell or high water. Thus *ex post* discretion for the firm is limited to the migrant employment decision. This contract form is intended to model situations where legal restrictions, trade unions, etc., in effect leave firms with little ability to lay off workers.

Native employment is therefore invariant across states, in the European mode, and equals the exogenously determined native labor force, when markets clear in equilibrium. Thus $e=1$ and N equals the supply of native workers.

I assume that chocolate firms are rational in the sense that, although they do not know in advance what the state of nature will be, they do understand sufficiently well the world in which they operate to know what would happen in each state. In particular, they know P_A and P_B and anticipate hiring migrants as described in Section I, Part A. Thus a_A and a_B are known to the firms.

Risk-neutral firms maximize expected profit, and this expected profit is driven to zero in equilibrium by the entry and exit of firms. This implies, in view of (3) and (4), that the equilibrium (state invariant) native wage equals the expected value of the marginal product of native labor: $(a_A + a_B)/2$.

In the alternative schema, which I call the *American mode*, firms and workers negotiate a state-dependent contract. This contract specifies the wage to be paid in each state to employed workers, and the conditional probability for each state that a worker under contract would not be laid off, should that state materialize. I assume for simplicity that the identities of laid-off workers are determined by a random draw from all contracted workers once the state of nature is known. Thus the probability of not being laid off should state i materialize equals state i employment as a fraction of total contracts.

Risk-averse workers prefer higher wages and higher probabilities of not being laid off —I assume that, unlike capitalists, workers cannot diversify risks away in the international capital market. Since chocolate firms compete with each other in the labor market, no firm would contract more native labor than it planned to employ in the more favorable state, B, since to do so would only make its contracts less attractive to workers. Thus the conditional probability of a contracted worker retaining a job is unity in state B and e in state A. Workers who are laid off from the chocolate industry are employed in a secondary sector which produces the composite traded good, using only native labor, at constant marginal productivity α.

Let $h(w)$ denote the strictly concave utility function of each native worker. If w_i denotes the native wage in state i, the contract (w_A, w_B, e) will yield a worker the expected utility

$$(6) \quad u(w_A, w_B, e) = \tfrac{1}{2}h(w_B) + \tfrac{1}{2}\left[eh(w_A) + (1-e)h(\alpha)\right].$$

In what follows I analyze in detail only the American mode. Modifications implied by the European mode are easily noted.

D. *Equilibrium in the Native Labor Market (American Mode)*

Each chocolate firm is competitive in the market for native workers, so the contract it offers must yield utility $u(w_A, w_B, e)$ equal to the market determined level u. Thus the decision problem faced by each chocolate firm is to choose (w_A, w_B, e) to maximize expected profit per contract:

$$\tfrac{1}{2}(a_A - w_A)e + \tfrac{1}{2}(a_B - w_B),$$

subject to $\quad u(w_A, w_B, e) = u; \quad e \leq 1.$

Because of constant returns to scale, the number of contracts offered by an individual

chocolate firm is indeterminate but immaterial.

The first-order conditions for this problem[5] imply a state-invariant wage, a basic result of ("first generation") implicit contract theory: $w_A = w_B \equiv w$. In market equilibrium, the expected profit of chocolate firms is driven to zero by entry and exit, and u also adjusts so as to clear the market for native labor. Thus N in (1)–(4) will equal the (assumed exogenous) supply of native laborers. The zero-profit condition plus the first-order conditions reduce to the following:

$$e(a_A - w) + (a_B - w) = 0 \tag{7}$$

$$a_A - w + \phi(w) \geq 0 \tag{8}$$

$$(1-e)[a_A - w + \phi(w)] = 0 \tag{9}$$

where $\phi(w) \equiv [h(w) - h(\alpha)]/h'(w)$, so that $\phi' > 1$ if $w > \alpha$. In addition, the utility of each native worker is given by (6).

E. *Full Equilibrium*

Consider in more detail the native labor market described by (6)–(9). If the economy remains completely specialized to chocolate in both states, that is, if $e = 1$, (9) will be met and (7) determines the wage

$$w = (a_A + a_B)/2. \tag{10}$$

Native labor is paid the expected value of its marginal product, as in the European mode. This will indeed be an equilibrium if the wage specified in (10) is consistent with (8).

Let the function $\omega(a_A)$ be the solution to $a_A - \omega + \phi(\omega) = 0$. The graph of this function is called the Risk Premium curve and is labeled *RP* in the third quadrant of Figure 1. The Risk Premium curve shows for each value of a_A what wage w is necessary to pay workers the premium $\phi(w)$ above the value of their marginal product, as compensation for the risk of unemployment they assume if e falls below 1. Note that $\omega' < 0$ and $\omega = \alpha$ if $a_A = \alpha$. Points on or below this curve satisfy (8).

The line labeled *EA* in Figure 1 graphs (10) and so shows, for each value of a_A, the wage that equals the expected value of the marginal product of native labor, given the state B equilibrium. This line intersects the w axis at $w = a_B/2$ and the 45° line at $w = a_B = a_A$. Points on this line below the *RP* curve satisfy (8) and so are consistent with equilibrium. Thus for any value of a_A at least as great as a_0 in Figure 1, the corresponding point on *EA* indicates the equilibrium wage, and $e = 1$.

If a_A is strictly less than a_0, it is not possible to have $e = 1$: the economy will not completely specialize in chocolate production in state A. In this case (9) will require $w = \omega(a_A)$ and (7) will then determine e:

$$e = [a_B - \omega]/[a_A - \omega]. \tag{11}$$

The Equilibrium Employment curve, labeled EE' in the second quadrant of Figure 1, records all combinations of w and e consistent with (11). Together with the linear segment $E'E''$, corresponding to full employment in both states, it displays all combinations of those two variables consistent with equilibrium in the native labor market.

It is now possible to determine, for each value of a_A, that level of e consistent with equilibrium in the market for native labor. This is done in Figure 1 and the result recorded in the first quadrant as the Native Labor schedule, labeled *NL*. Full equilibrium is determined by the intersection of the International Market and Native Labor schedules.

Figure 1 depicts an equilibrium in which Switzerland is incompletely specialized in state A. The values of a_A, w, and e can be read off the diagram, and the (*ex ante*) utility of native residents is then given by (6). The curve labeled u in the second quadrant is the indifference curve corresponding to the equilibrium utility level. Indifference curves must cut the EE' curve from above, as drawn.[6]

[5]See the Appendix for further details.

[6]In the European mode, the Native Labor schedule is a horizontal line with $e = 1$, point E' is directly above $w = a_B/2$, and the Equilibrium Employment and Risk Premium curves become irrelevant.

II. Characteristics of Equilibrium

The degree of the Swiss (marginal) comparative advantage in chocolate in state i can be thought of as measured by $a_i - \alpha$. By assumption there is such an advantage in state B. If the equilibrium $a_A > \alpha$ as well, Switzerland always has a comparative advantage in chocolate and always specializes completely in the product. In this case, the sole significance of the quasi-permanent relations between firms and native workers is stabilization of the native wage: if instead natives, like migrants, were hired in spot markets after realization of the state of nature, the wage would equal a_i but Switzerland would still specialize completely in chocolate production in both states.

When $a_A < \alpha$, Switzerland has a (marginal) comparative disadvantage in chocolate in state A, and the native contractual arrangements influence the pattern of specialization. For example, if $a_0 \leq a_A < \alpha$, the Swiss still specialize completely in chocolate production, whereas spot markets would cause them to diversify.

The Swiss populace is exposed to uncertainty through both of the international markets in which it deals. Risk-neutral firms provide partial insulation from this uncertainty. The insulation they can provide is in turn related in two ways to their own activity on world markets. By hiring foreign workers in good times and discharging them in bad times, firms help insulate Swiss workers. The second method has to do with the pricing of chocolate. Equation (3) says that in state A firms pay migrants the value of their marginal product, and (8) implies that Swiss workers are paid more than the value of their marginal product. Thus Switzerland exports chocolate at a price less than its cost of production, that is, Swiss products are dumped[7] in world markets in state A. Note that this dumping would appear aggressive to foreign competitors, since Swiss firms employ no fixed factors at all but dump in order to avoid laying off workers. Chocolate is sold above cost in state B, so that firms expect to break even in the long run.[8]

The expulsion of foreign workers and the dumping of commodities has each received great attention throughout the previous decade, but the relation between the two has been largely ignored. The present paper offers a framework for the investigation of this relationship.

The Swiss economy is exposed to uncertainty through the two international markets in which it deals. The stochastic process relating to labor migration—reflected in the magnitudes of v_A and v_B—would in practice be quite sensitive to national immigration policy. For example, a guest worker system might be operated in a relatively closed fashion, allowing only a modest volume of immigration and adjusting that in response to disturbances so as to stabilize the migrant wage faced by individual domestic firms. Alternatively, the authorities might be more willing to allow the domestic economy to exploit the opportunity to obtain cheap migrant labor services and so operate a more open guest worker system, insisting that all migrants have jobs but not trying to shield chocolate firms from external fluctuations in

[7]See my 1982 article for a fuller, two-country theory of dumping.

[8]At this point, consider briefly Switzerland's international payments. In state A the excess of the value of Swiss chocolate exports over imports and remittances of migrant wages equals

$$eN[P_A f(m_A) - w - v_A m_A] = eN(a_A - w).$$

(Note that this is independent of how migrants divide their wages between remittances and local expenditures.) Let π_A denote total state A profits of the chocolate industry, γ the share of that industry owned by Swiss capitalists, and π_A^F total profits received by Swiss capitalists from ownership of industries in the rest of the world. Thus the state A Swiss current account surplus equals

$$eN[a_A - w] + \pi_A^F - (1-\gamma)\pi_A.$$

But $eN(a_A - w) = \pi_A$, so the current account surplus equals $\pi_A^F + \gamma\pi_A$, or total state A profits received by Switzerland. In like manner, the current account surplus (identically equal to net Swiss capital accumulation) in state B equals total profit in that state. Assuming that Swiss capitalists have been able to diversify away all risk by means of the international exchange of securities, the current account is state invariant.

migrant wages. The authorities might also vary their administration of a guest worker system in response to fluctuations in the world demand for chocolate. This diversity of possibilities demands an investigation of how economic performance varies in response to changes in the stochastic process facing Switzerland. I turn next to this problem.

III. Alternative Stochastic Processes

Consider now how the equilibrium described in Section I is influenced by changes in the stochastic process confronting Switzerland. Note that this comparative statics exercise is to be distinguished from the simpler question—whose answer is immediate from the foregoing analysis—of how state A behavior differs from state B behavior for a given stochastic structure. There are two reasons for addressing the present question. The first is the possibility, raised above, that the stochastic structure is sensitive to immigration policy. The second is that my assumption of just two states of nature, made solely for expositional purposes, is too limiting in the presence of two random variables. For example, one might wish to ask how Swiss behavior depends upon whether v and X are positively correlated or negatively correlated. Thus a full understanding of the model requires consideration of alternative stochastic processes.

Changes in X_B and v_B produce changes in a_B. The effects, easily traced through Figure 1, imply shifts of the Native Labor schedule and consequent movements along the International Markets schedule. Changes in X_A and v_A by contrast shift IM and thereby causes movements along NL. The degree of volatility confronting Switzerland is determined by the differences $X_B - X_A$ and $v_B - v_A$. To have a frame of reference, I shall take X_B and v_B as immutable (so that equilibrium values of P_B, m_B, and a_B do not change) and consider alternative values of X_A and v_A. Each pair of values of X_A, v_A determines, from Section I, equilibrium levels of e, P_A, m_A, a_A, and w. I shall examine in turn the effects of variations in X_A and v_A upon equilibrium.

A. *World Demand for Chocolate*

Consider, first, a rise in X_A in the proportion $\hat{X}_A$. With unchanged X_B, this constitutes an increase in expected world demand for chocolate. I have labeled the states so that $a_A < a_B$, so there is a presumption (but not a certainty) that $X_A < X_B$. If so, the rise in X_A constitutes also a reduction in the volatility of world chocolate demand. As discussed in Section II, Part B above, the International Markets schedule in Figure 1 shifts upward in the proportion $\hat{X}_A$. Thus equilibrium moves to the right along the Native Labor schedule. If initially $e=1$, this produces a rise in w and consequent rise in utility u. If $e<1$ initially, w *falls* but e rises, and on balance the increase in job security outweighs the reduction in the wage rate in the sense that u still rises. Note the contrast in *ex post* consequences in the two cases. With complete specialization ($e=1$) the *ex ante* and *ex post* utility of each Swiss resident is $h(w)$ and clearly rises as a consequence of the rise in expected world demand for chocolate. If $e<1$, the *ex ante* utility u of each Swiss likewise increases. But should state B materialize, the *ex post* utility $h(w)$ of each resident will be *lower* than what it would have been with the lower X_A. Should state A materialize, the *ex post* utility of laid-off workers, $h(\alpha)$, is of course unchanged, and the *ex post* utility of chocolate workers who would have been employed there even if X_A had not risen (workers' identities are of course unknown) is strictly lower. The only *ex post* utility that is increased[9] is that of chocolate workers in state A who would have been laid off with the lower X_A.

B. *Migrant Wages*

Now consider the effects of a reduction in state A migrant wages in the proportion $\hat{v}_A < 0$. As discussed in Section II, Part B above, the International Markets schedule shifts up or down, according as σ_A is less than η_A or not, and equilibrium moves in a corresponding fashion along the Native Labor schedule.

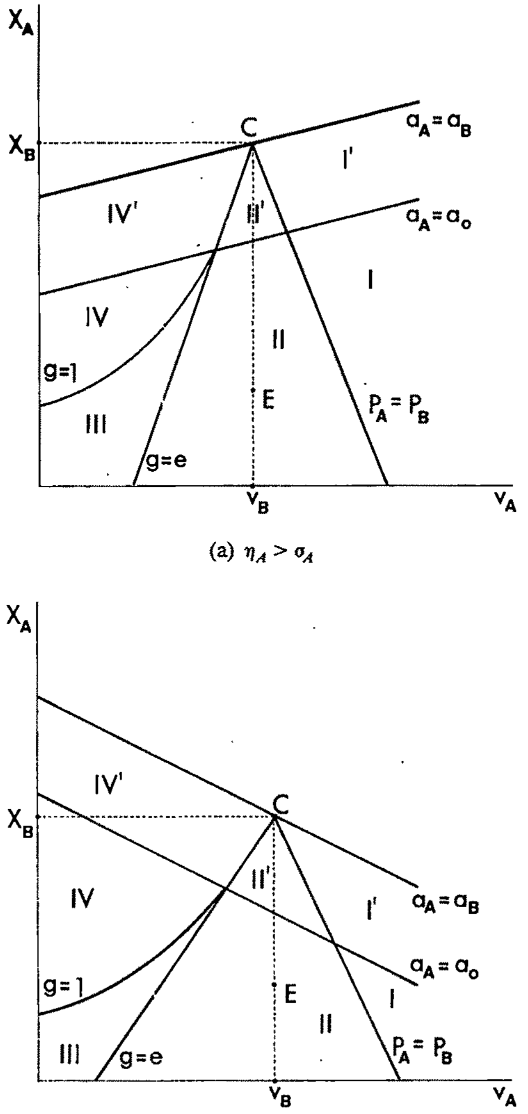
[9]Note the analogy with the Stolper-Samuelson theorem in factor-endowments models.

Note, incidentally, that we can easily see, with the aid of Figure 1, that a reduction in v_A brings about the paradoxical result that native utility falls, as long as $\sigma_A > \eta_A$. (It can easily be shown that a fall in both v_A and v_B, so that migrant wages fall with certainty, can also reduce u.) But the reason is simple: lower migrant wages can induce chocolate firms to increase production by enough so that the resulting deterioration in the terms of trade produces a welfare loss.[10] Such possibilities disappear if Switzerland becomes small in the world chocolate market (so that η_A becomes very large), or if the country levies an optimal export tax on chocolate.

C. *The Array of Possibilities*

It is time to examine all possible equilibria corresponding to different stochastic processes. With X_B and v_B immutable, each pair of values of X_A and v_A determines a particular equilibrium. Each point in each panel of Figure 2 corresponds to a distinct hypothetical stochastic process confronting Switzerland. At point C, $X_A = X_B$ and $v_A = v_B$, so that the future is certain. Thus both states have identical, complete-specialization equilibria: $a_A = a_B$ and $e = 1$. Employment of migrants also does not vary across states, so that $g = 1$, where g denotes the ratio of state A employment of migrants to state B employment, and thus $g = em_A/m_B$. At C the present model in effect reduces to the standard factor-endowments model with complete specialization in the production of exportables.

Consider first combinations of X_A and v_A that leave a_A unchanged. It follows from Parts A and B above that, in order that a_A not change, increases in X_A must be accompanied by increases in v_A if $\eta_A > \sigma_A$ and by decreases in v_A if $\eta_A < \sigma_A$. Each case is allotted its panel in Figure 2, and in each panel two a_A contours are drawn. Downward movements bring one to contours corresponding to lower values of a_A. The $a_A = a_B$ contour must pass through the certainty point C; points above this line will be ignored



(a) $\eta_A > \sigma_A$

(b) $\eta_A < \sigma_A$

FIGURE 2

because they correspond to points below with the roles of the states reversed.

If is evident from Figure 1 that constant values of a_A imply constant values of e, w, and *ex ante* utility u as well. Thus along the $a_A = a_B$ line the Swiss enjoy the same utility, both *ex ante* and *ex post*, as in the certainty case. The $a_A = a_0$ line corresponds to the lowest value of a_A for which $e = 1$, and to the lowest wage w. Moving from $a_A = a_B$ to

[10]See Bhagwati (1968 and 1973) for the appropriate analogous treatment of immiserizing growth.

TABLE 1—POSSIBLE EQUILIBRIA

Zone	Employment	Terms of Trade
I	$1>e>g$	$P_A>P_B$
I'	$1=e>g$	$P_A>P_B$
II	$1>e>g$	$P_A<P_B$
II'	$1=e>g$	$P_A<P_B$
III	$1>g>e$	$P_A<P_B$
IV	$g>1>e$	$P_A<P_B$
IV'	$g>1=e$	$P_A<P_B$

$a_A=a_0$ reduces a_A and w, keeps Switzerland completely specialized in both states ($e=1$), and reduces utility (*ex ante = ex post*). Continuing downwards brings one to equilibria with lower values of e and a_A, higher values of w, and lower *ex ante* utility (but higher *ex post* utility for chocolate workers).

The line labeled $P_A=P_B$ shows the assumptions that result in sticky terms of trade across states. It slopes downwards because P_B is constant and P_A is increased by rises in both X_A and v_A. $P_A>P_B$ above this line and $P_A<P_B$ below.

Swiss employment in chocolate varies across states in the same proportion as migrant employment on the $g=e$ line. This happens when $m_A=m_B$, which, in view of the constancy of m_B, requires a constant m_A. Equation (3) implies that m_A will be unaltered if increases in v_A are accompanied by large enough increases in X_A to raise P_A in the same proportion (thereby raising a_A as well). Thus the $g=e$ line has a positive slope and cuts a_A contours from below. Above this line $g>e$ and below it $g<e$.

Finally, migrant employment is stable across states along the $g=1$ schedule. This line is found by solving the system of equilibrium conditions, (1), (3), (7), and (8)—with equality—for X_A as a function of v_A, subject to the restriction that $em_A=m_B$. Points above this line imply $g>1$ and points below it imply $g<1$. The various lines divide each panel of Figure 2 into several zones, whose properties are summarized in Table 1.

With the European mode, the analysis is basically the same except that only the full-employment states (*I'*, *II'*, and *IV'*) are relevant (so of course there can be no case where migrant employment is less volatile than native). Sufficiently low values of X_A (and/or of X_B) cause the economy to instead specialize completely in the production of the composite traded consumption good in both states.

D. *Dumping Goods and Dumping Migrants*

Suppose initially that $v_A=v_B$, as at points E in Figure 2. Then Switzerland is exposed to uncertainty solely through the chocolate market. As is clear from the figure, this necessarily results in an equilibrium in zone *II* or *II'*. Suppose the former—whenever the unfavorable state A materializes, Swiss workers are laid off ($e<1$). But Swiss firms moderate this in two ways: they reduce prices ($P_A<P_B$) and dump chocolate abroad, and they also lay off migrant workers in greater proportion than Swiss workers ($g<e$). Thus both chocolate and migrants are dumped abroad when foreign demand for Swiss goods falls. Zone *II* equilibria would characterize guest worker systems administered in a "closed" way as discussed in Section II above.

Now let v_A fall; that is, consider leftward movements in Figure 2. This consists of cases where low values of X_i are accompanied by low values of v_i: unfavorable commodity-market conditions are correlated with favorable labor-market conditions. This would be the case if Swiss chocolate, and its close substitutes, were to a large degree sold in the same countries from which migrant labor comes: both conditions would be sensitive to aggregate demand in those countries. Such movements depart further from the $P_A=P_B$ line in Figure 2, implying larger state A price reductions in chocolate to stimulate demand. This is made possible by the cost-reducing effect of lower values of v_A. The latter also furnishes a motive to reduce fluctuations in the employment of migrants, and g indeed rises with leftward movements in Figure 2. Thus domestic employment is tugged in two directions by the tendency, on the one hand, to *substitute* foreign workers for domestic workers in state A—made easier by higher values of σ_A—and by the tendency, on the other, to increase state A production—made

easier by higher values of η_A—which affects foreign and migrant employment in a *complementary* way. If Switzerland is basically a small country in the world chocolate market (very large η_A) the complementarity aspect necessarily dominates, but with less than infinite elasticity of demand the opportunity to expand chocolate sales is limited. In general, the complementarity or the substitutability aspect will dominate according as $\eta_A > \sigma_A$ or the reverse, since Figure 2 shows that this is the condition deciding whether a leftward movement increases or decreases a_A, e (if initially less than unity), and *ex ante* utility u.

If v_A falls enough the economy finds itself in zone *III* where migrant employment is more stable than domestic employment. But migrants are still being used to stabilize Swiss employment conditions: since migrant wages fall when the demand for chocolate falls, retaining more foreign workers on the payroll in state A reduces costs and may still allow more Swiss workers to be retained also, if $\eta_A > \sigma_A$. The problem is that maintaining employment through maintaining production in state A causes the terms of trade to deteriorate: prices fall as well as costs. If migrants and natives are not sufficiently complementary and/or if chocolate demand is not sufficiently elastic, $\sigma_A > \eta_A$ and *ex ante* utility falls.

If chocolate conditions and foreign labor conditions are sufficiently negatively correlated, the economy will be in zone *IV*, so that employment of migrants actually rises in state A. Chocolate dumping and migrant dumping cease to be complementary and become substitutes: chocolate is dumped in state A and migrants in state B. It could also become unclear whether the migrant labor market is being exploited to shield Switzerland from shocks in the chocolate market, or whether the reverse is the case. For example, if $\eta_A < \sigma_A$ zone *IV* (or *IV′*) contains cases where $X_A > X_B$ and $v_A < v_B$: in state A domestic unemployment due to the substitution of cheap migrant labor is ameliorated by an increased foreign demand for chocolate. In any event the dumping of chocolate that takes place in state A would seem very aggressive indeed to Switzerland's foreign competitors: the Swiss *increase* their employment of migrants to be able to dump chocolate abroad!

Positions to the right of point E in Figure 2 correspond to cases where favorable commodity conditions are positively correlated with favorable (for the Swiss) migrant labor conditions. This could be the case if chocolate is sold in a part of the world with economic conditions negatively correlated with those in the part of the world from which the migrants come. Since labor costs rise in state A, the Swiss are less able to reduce prices, though they still dump chocolate. Instead they are more inclined to lay off migrants and substitute domestic labor: $e > g$ to the right of E. If $\sigma_A > \eta_A$, this substitution effect will outweigh the complementarity effect resulting from the depressing influence of higher costs, and the increase in v_A will move the Swiss closer to the $e = 1$ region. With high enough positive correlation between chocolate and migrant labor conditions. Switzerland will be in zone *I*, characterized by the apparent paradox that the Swiss increase the price of chocolate when they dump it ($P_A > P_B$). The reason is simply that they raise prices to cover part, but not all, of their increased labor costs.

Public discussion of labor migration emphasizes two, basically inconsistent, alleged benefits to host countries: isolation of native workers from the business cycle by varying the employment of migrants (which requires a substitutability of domestic and foreign labor in production), and preservation of domestic jobs in domestic industries by reliance on cheap migrant labor (which emphasizes complementarity). The present analysis provides a framework for the study of these two alleged benefits and shows how they relate to the nature of the stochastic process facing the host country, as well as to basic parameters such as elasticities of demand and of substitution (which determine the shape and position of the curves in Figure 2). The dividing line between the two cases is the $g = e$ curve, the locus of possibilities for which production techniques are unchanged across states. In zones *I* and *II*, the Swiss respond to a recession by substituting domestic workers for migrants and sending

the latter home: the recession is (partly) passed on to foreign workers. In zones *III* and *IV*, a recession by contrast causes migrants to be substituted for domestic labor, and it is the resulting cost reduction that allows larger chocolate production—and native employment—than would otherwise be possible. This latter possibility requires, but is not assured by, a negative correlation between conditions in the two markets. A higher elasticity of chocolate demand increases the ability of a price reduction—and thus cost reduction—to maintain chocolate output. Thus larger values of η_A cause the $g = e$ line to become steeper (this is confirmed in the Appendix) and so expand zones *III* and *IV* at the expense of *I* and *II*. But, unless η_A is infinite, zone *II* necessarily contains cases where $v_A < v_B$.

The parameter σ_A measures the extent to which domestic and migrant labor are similar and thus highly substitutable for each other. As σ_A increases, zone *IV* expands at the expense of the other three zones.

IV. The Effect of Migration and Migration Policy

The preceding discussion of alternative stochastic processes was intended to expose the role of temporary migration in the host economy. I turn now to its significance: in what ways is the behavior of the host country different from what it would be if there were no temporary migration?

A. *Captive Migrants*

Suppose, first, that the economy possesses a fixed supply M of unskilled labor that does not vary in response to the state of nature. The supply M might consist of native unskilled workers or permanent immigrants. Unskilled labor is hired in spot markets after the state of nature is known. Thus, in equations (1)–(4), $m_B = M/N$, $m_A = M/eN$, and v_A and v_B are now endogenously determined. In other respects the model is unchanged. A distinct equilibrium is now determined for each set of values of X_A, X_B, and M rather than, as before, for each set of values of X_A, X_B, v_A, and v_B.

Each combination of X_A, X_B, and M will give an equilibrium corresponding to some point on the $g = 1$ line of the relevant version of Figure 2. Thus, in terms of the classification in Table 1, equilibrium will be on the boundary of zones *II'* and *IV'* or on that of *III* and *IV*. Equilibria with the characteristics of all other zones are thus essentially dependent upon labor mobility. For example, there will no longer be cases where, in state *A*, chocolate firms simultaneously raise prices and dump their product abroad.

Next, I wish to compare an equilibrium with labor mobility to a "naturally corresponding" one without. But what set of values of X_A, X_B, and M corresponds naturally to a particular choice of X_A, X_B, v_A, and v_B? One would wish, I should think, to keep X_A and X_B the same as with labor mobility, but in the choice of M one cannot avoid being arbitrary. Purely for the sake of illustration, then, I will arbitrarily choose that value of M that causes v_B to have the same value in the no-mobility case as in the comparison equilibrium with labor mobility.

Suppose, then, that initially $X_A = X_B$ and $v_A = v_B$ with migration possible. If migration is now ruled out, X_A and X_B are kept unchanged, and M is set to generate the same value of v_B as before, equilibrium will not in fact change: both cases correspond to point C in each panel of Figure 2. Now consider how equilibrium (migration and no-migration) responds to a comparative statics reduction in X_A. The change in the migration equilibrium corresponds to a downward vertical movement in each panel of Figure 2, from point C toward E (and perhaps beyond). The related no-migration equilibrium shifts down by the same vertical amount, but along the $g = 1$ curve instead.

Three conclusions are immediately apparent from Figure 2. First, both P_A and v_A are lower in the no-migration case: migration reduces the variability of prices across states. Second, the comparative welfare of native workers depends critically upon the relative magnitudes of η_A and σ_A. In panel (a), where $\eta_A > \sigma_A$, the presence of migration results in an equilibrium with a higher value of a_A. Thus, as a result of migration, native labor has a lower level of (*ex ante*) utility, in some

situations migration leads to state A unemployment when the absence of migration would preserve full employment in both states, and, if $e<1$ in both equilibria, it is lower with migration. Panel (b) shows that these conclusions are exactly reversed if $\eta_A < \sigma_A$. The reason for this discrepancy should be clear from the discussion in preceding sections. If migrants are not sent packing when the world demand for chocolate falls, production will be higher than it would otherwise be. If world demand for chocolate is sufficiently inelastic, the resulting deterioration in the terms of trade will be the dominant consideration. But if the elasticity of substitution is sufficiently low, the resulting reduction in the spot wage of unskilled workers will prove decisive.

The third observation is that v_A is always less without migration. Since v_B is the same in both cases, unskilled workers are a "captive" labor force in the sense that they unambiguously suffer because of the lack of the option to emigrate. Of course, a judgment about host-country welfare should be sensitive to that of unskilled workers as well as the skilled, once emigration is ruled out. On this basis, the absence of labor mobility harms both skilled workers and the unskilled if $\sigma_A > \eta_A$, but benefits the former at the expense of the latter when $\eta_A > \sigma_A$.

B. *Immigration Quotas*

I pointed out above that the choice of which size of M to compare to any v_A, v_B pair was inherently arbitrary. My choice involved a larger unskilled labor force in one state than migration would produce. We could of course consider other choices, in particular some that involve less unskilled labor in both states than would migration. I leave to the reader the straightforward investigation of alternative choices and I instead approach the question somewhat differently. Consider now the effects on the initial equilibrium of a quota on immigration.

Suppose a (state-invariant) quota Q. If Q is initially greater than the immigration that would originally occur in either state, it would of course have no effect, so suppose that Q is reduced until it just becomes binding in one state. If the initial equilibrium is in zone IV or IV' in Figure 2, this will be state B; otherwise it will be state A. Assume for now the latter.

Such a quota will not impinge upon equations (1) and (3), so the International Markets schedule in Figure 1 is unchanged.

The quota will reduce m_B in equation (2), causing P_B to rise in an amount inversely proportional to η_B. Thus a_B can either rise or fall. Not surprisingly, this depends upon whether η_B exceeds σ_B or not. Suppose it does, so that the effect of the quota is to lower a_B. (Equation (4) shows the effect of the quota on the state B migrant wage, which is accordingly driven above v_B. I suppose that these rents go to those migrants who happen to be fortunate enough to get in.)

It is immediate from the discussion of the full equilibrium in Section II, Part E, that the reduction in a_B shifts the Equilibrium Employment curve EE' in Figure 1 down, leaves the Risk Premium curve RP unchanged, and shifts the EA line to the left. These imply that that Native Labor schedule NL, in the first quadrant, shifts to the right.

If the initial equilibrium in the first quadrant of Figure 1 is characterized by $e=1$, a small shift of the NL schedule will not alter that fact. But the right-hand shift of the EA line implies that w falls. The prospect of a binding immigrant quota should state B materialize results in a lower wage for (fully employed) native workers in both states: firms will pay less for native workers because the prospect of a constraint on migrant employment in one state reduces the expected value of the marginal product of native labor. Native workers are unambiguously worse off, *ex ante* and *ex post*.

If instead the initial equilibrium features $e<1$, imposition of a quota causes that equilibrium to shift down and to the right along the International Markets schedule. Thus e falls and a_A rises, and the fixity of the RP curve implies that w again falls. The prospect of a binding quota in state B again reduces the *ex ante* value of native employment. This is reflected again by a reduction of w (in both states), but also now by a reduction in e as well. Because of the prospect of a quota on migrants, firms are less

willing to commit themselves to natives, and so plan to rely relatively *more* on migrants should the nonbinding state A arise. It is conceivable that the imposition of a quota could actually increase *ex post* immigration! In any event, native workers are unambiguously worse off, both *ex ante* and *ex post*.

If $\sigma_B > \eta_B$, the imposition of a quota causes a_B to rise and all the above conclusions are reversed. The reasons for this should be clear by now.

If the initial equilibrium is in zone IV or IV', so that a small quota is initially binding in state A rather than B, the effects are quite different. Now equations (2) and (4) are unaffected, so that a_B does not change. Instead it follows from (1), and the definition of a_A, that a small quota will shift the International Markets schedule to the left or the right, according as η_A exceeds σ_A or not. Suppose the former. It is immediate from Figure 1 that, if the initial equilibrium features full employment $e = 1$, the effect of the quota is to lower w in both states, just as in the preceding case. If initially $e < 1$, the leftward shift of the International Markets schedule produces a downward movement along the Native Labor schedule and a corresponding downward movement along the EE' curve. The quota lowers e but raises w. As the economy moves to a lower indifference curve, the *ex ante* utility of native workers necessarily declines. But should state B actually materialize, the fact that a quota would have been binding in state A makes native workers better off *ex post*! If instead state A comes about, there is an *ex post* distributional effect. Native workers who retain their jobs are better off while those who lose their jobs, as a result of the quota, are worse off. If $\sigma_A > \eta_A$, all these conclusions are reversed.

To summarize, the effects of an immigration quota depend upon three considerations: whether $\eta_A > \sigma_A$ or not; whether there is initially full employment in both states; and whether migrant dumping is a complement to or a substitute for commodity dumping. If $e = 1$ initially, a quota necessarily lowers the native wage in both states if $\eta_A > \sigma_A$ and raises it if $\eta_A < \sigma_A$. If initially $e < 1$, the quota lowers e and *ex ante* welfare if $\eta_A > \sigma_A$ (and raises them in the opposite case), but other *ex post* consequences depend critically upon whether or not migrant dumping substitutes for commodity dumping. More stringent quotas can be analyzed in similar fashion. When the quota becomes binding in both states, so that there is a state-invariant unskilled labor force $M = Q$, the International Markets and Native Labor schedules shift simultaneously.

V. Concluding Remarks

This paper has developed a theory of international trade and labor mobility built squarely upon the five stylized facts pointed to at the outset, but which reduces to a standard factor-endowments model under special circumstances (certainty). The dominant theme is that of the smoothing out of native labor employment fluctuations by means of a combination of commodity dumping and migrant dumping. It is the link between these two that in fact clarifies the seemingly inconsistent claims that the substitution of domestic labor for migrants allows unemployment to be exported and that the complementarity of the two types of labor permits cost reductions that sustain domestic employment. Three features emerged as key determinants of behavior: the elasticity of substitution between migrant workers and natives in production; the world elasticity of demand for home exports; and the correlation between conditions in the export market and in the migrant labor market. The role of these determinants, the relations between production, labor markets and international trade, and the implications for host-country welfare were investigated in detail.

The present paper leans heavily on simplifying assumptions, because my prime concern has been to sketch out the essentials of a theory radically different from the standard one derived from factor-endowment trade models.[11] The need for further development

[11] The present theory is actually closer in spirit to the dual economy models employed in development economics. See Corden (1974, ch. 6). Of the references cited

should be obvious. Potential developments of greatest interest, in my judgement, include consideration of the effects of uncertainty of domestic origin and also of the relation between migration and internal resource allocation and income distribution. It is often claimed that the ability to hire migrants reduces the need to reallocate native labor between sectors or between regions in response to changing circumstances. Analysis along these lines would entail the development of a multisectoral version of the present model. It is often noted that migrant workers are much more intensively employed in some sectors than in others. The relation of this observation to the reallocation argument is unclear and could certainly benefit from formal analysis.[12]

The present analysis was solely from a host-country perspective. Analysis of source countries is necessary to obtain a full treatment of international equilibrium and to investigate the welfare of migrants and of those they leave behind. Also of considerable interest would be the investigation of the consequences of asymmetric access to information by domestic firms, native workers, and migrant laborers.

APPENDIX

The firm's problem is to choose w_A, w_B, e to maximize

$$\text{(A1)} \qquad \tfrac{1}{2}(a_A - w_A)e + \tfrac{1}{2}(a_B - w_B),$$

subject to

$$\text{(A2)} \qquad \tfrac{1}{2}h(w_B) + \tfrac{1}{2}[eh(w_A) + (1-e)h(\alpha)] = u$$

$$\text{(A3)} \qquad e \leq 1.$$

The first-order conditions for this problem are

$$\text{(A4)} \qquad \frac{e}{2}[\lambda h'(w_A) - 1] = 0$$

$$\text{(A5)} \qquad \tfrac{1}{2}[\lambda h'(w_B) - 1] = 0$$

$$\text{(A6)} \qquad \tfrac{1}{2}[a_A - w_A + \lambda h(w_A) - \lambda h(\alpha)] - \mu = 0$$

$$\text{(A7)} \qquad \mu(1-e) = 0$$

$$\text{(A8)} \qquad \mu \geq 0,$$

where λ and μ are the dual variables. Conditions (A4) and (A5) imply that $w_A = w_B \equiv w$, and $\lambda = 1/h'(w)$. Substituting this value of λ into (A6) and noting (A8) yields equation (8) in the text. Multiplying (A6) by $(1-e)$ and noting (A7) yields equation (9). In equilibrium, u will adjust to clear the native labor market, and firms will enter and exit to drive expression (A1) to zero. This latter condition is equation (7), and now (A2) merely measures the equilibrium value of u.

Differentiation of the definitions of a_A and a_B and of equations (1)–(4), (7) and (8) (with equality) yield the following (for $e < 1$):

$$\text{(A9)} \qquad \sigma_A(\hat{a}_A - \hat{P}_A) = \theta_{MA}\hat{m}_A$$

$$\text{(A10)} \qquad \sigma_B(\hat{a}_B - \hat{P}_B) = \theta_{MB}\hat{m}_B$$

$$\text{(A11)} \qquad \hat{X}_A - \eta_A\hat{P}_A = \hat{e} + \hat{N} + \theta_{MA}\hat{m}_A$$

$$\text{(A12)} \qquad \hat{X}_B - \eta_B\hat{P}_B = \hat{N} + \theta_{MB}\hat{m}_B$$

$$\text{(A13)} \qquad \theta_{DA}\hat{m}_A = \sigma_A(\hat{P}_A - \hat{v}_A)$$

$$\text{(A14)} \qquad \theta_{DB}\hat{m}_B = \sigma_B(\hat{P}_B - \hat{v}_B)$$

$$\text{(A15)} \qquad \frac{\theta_{DA}\theta_{MA}}{\sigma_A}\hat{m}_A = \gamma_A\hat{w} - \theta_{DA}\hat{P}_A$$

$$\text{(A16)} \qquad \frac{\theta_{DB}\theta_{MB}}{\sigma_B}\hat{m}_B = \gamma_B\hat{w} - \theta_{DB}\hat{P}_B + (\theta_{DB} - \mu_{DB})\hat{e}.$$

In these equations, $\theta_{Di} \equiv 1 - \theta_{Mi}$, $\mu_{Di} \equiv w/P_i f(m_i)$, $\lambda \equiv w\phi'(w)/\phi(w)$, and $\gamma_i \equiv \lambda\theta_{Di} + (1-\lambda)\mu_{Di}$ for $i = A, B$.

---

in ft. 2, Ohlin actually comes closest to the concerns of the present paper.

[12] Such a multisectoral model would also allow more extensive formal utilization of the ideas in the naturally related area of segmented labor markets. See Peter Doeringer and Piore (1971) and David Gordon, Richard Edwards, and Michael Reich (1982).

To derive the properties of the various curves in Figure 2, set $\hat{X}_B = \hat{v}_B = \hat{N} = 0$. Then

$$(a_A = a_0;\ \hat{a}_A = 0)\,\hat{X}_A = (\eta_A - \sigma_A)\theta_{MA}\hat{v}_A,$$

$$(P_A = P_B;\ \hat{P}_A = 0)\,\hat{X}_A$$

$$= -\theta_{MA}\left[\frac{\sigma_A}{\theta_{DA}} - \frac{\gamma_B}{\gamma_A}\frac{1}{\theta_{DB} - \mu_{DB}}\right]\hat{v}_A,$$

$$(g = e;\ \hat{m}_A = 0)\,\hat{X}_A$$

$$= \left[n_A - \frac{\gamma_B}{\gamma_A}\frac{\theta_{DA}}{\theta_{DB} - \mu_{DB}}\right]\hat{v}_A,$$

$$(g = 1;\ \hat{e} + \hat{m}_A = 0)\,\hat{X}_A$$

$$= \left[\eta_A - \frac{\gamma_B}{\gamma_A}\frac{\sigma_A - \eta_A}{\dfrac{\sigma_A}{\theta_{DA}} - \dfrac{\gamma_B}{\gamma_A}\dfrac{1}{\theta_{DB} - \mu_{DB}}}\right]\hat{v}_A.$$

Thus the slope of the $e = 1$ line depends upon the sign of the term $\eta_A - \sigma_A$, the $P_A = P_B$ line has negative slope, and the $g = e$ and $g = 1$ lines both have positive slopes, with the latter the steeper. In the special case $\eta_A = \sigma_A$, the a_A lines are flat and the elasticity of $g = 1$ is the elasticity of demand.

As a second comparative statics exercise, consider a general rise in migrant wages: $\hat{v}_A = \hat{v}_B = \hat{v}$. Let $\hat{X}_A = \hat{X}_B = \hat{N} = 0$ in (A9)–(A16) and solve for the endogenous variables. This gives the following

$$\text{(A17)}\quad \hat{P}_A = \frac{\theta_{DA}\gamma_A}{Z}\left[(\theta_{DB} - \mu_{DB})\theta_{MA}(\sigma_A - \eta_A) - \frac{\theta_{DB}\theta_{MB}}{\rho_A}(\sigma_B - \eta_B) + \frac{Z}{\theta_{DA}\gamma_A}\right]\hat{v}$$

$$\text{(A18)}\quad \hat{P}_B = (\theta_{MB}\sigma_B/\rho_B)\hat{v}$$

$$\text{(A19)}\quad \hat{M} = -(\eta_B\sigma_B/\rho_B)\hat{v}$$

$$\text{(A20)}\quad \hat{e} = (1/Z\rho_B)[\theta_{MB}\theta_{DB}\gamma_A\rho_A(\sigma_B - \nu_B) - \theta_{MA}\theta_{DA}\gamma_B\rho_B(\sigma_A - \eta_A)]\hat{v}$$

$$\text{(A21)}\quad \hat{w} = (1/Z\rho_B)[\theta_{DA}\theta_{MA}(\theta_{DB} - \mu_{DB})\rho_B(\sigma_A - \eta_A) - \theta_{DB}\theta_{MB}\theta_{DA}(\sigma_B - \eta_B)]\hat{v}$$

$$\text{(A22)}\quad \hat{g} - \hat{e} = (1/Z\rho_B) \times [\theta_{MA}(\theta_{DB} - \mu_{DB})\gamma_A\sigma_A\rho_B(\sigma_A - \eta_A) - \theta_{DB}\theta_{MB}\gamma_A\sigma_A(\sigma_B - \eta_B) + Z(\eta_B\sigma_B - \rho_B\sigma_A)]\hat{v}$$

where M denotes the employment of migrants in state B, and $Z = \rho_A(\theta_{DB} - \mu_{DB})\gamma_A - \theta_{DA}\gamma_B < 0$. If $\hat{v} > 0$, then P_B increases and M falls. The direction of change of all other terms is ambiguous, with e rising or falling as σ_A exceeds or falls short of η_A; the other terms are all sensitive to the sizes of these elasticities as well as to various shares. In the special case $\eta_A = \sigma_A = \sigma_B = \eta_B$, these equations give

$$\hat{M} < 0;\quad \hat{P}_A - \hat{v} > \theta_{MB}\hat{v} = \hat{P}_B > 0;$$
$$\hat{e} = \hat{g} = \hat{w} = 0.$$

REFERENCES

Azariadis, Costas, "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, December 1975, *83*, 1183–202.

Bhagwati, Jagdish N., "Distortions and Immiserizing Growth: A Generalization," *Review of Economic Studies*, October 1968, *35*, 481–85.

______, "The Theory of Immiserizing Growth: Further Applications," in Michael B. Connelly and Alexander D. Swoboda, eds., *International Trade and Money*, Toronto: University of Toronto Press, 1973.

______, "International Factor Movements and National Advantage," 9th V. K. Ramaswami Memorial Lecture, Delhi, 1979.

______ **and Brecher, Richard A.,** "National Welfare in an Open Economy in the Presence of Foreign-Owned Factors of Production," *Journal of International Economics*, February 1980, *10*, 103–16.

Birks, J. S. and Sinclair, C. A., *International Migration and Development in the Arab Region*, Geneva: International Labor Organization, 1980.

Brecher, Richard A. and Bhagwati, Jagdish N., "Foreign Ownership and the Theory of Trade and Welfare," *Journal of Political Economy*, June 1981, *89*, 497–511.

Corden, W. Max, "The Economic Limits to Population Increase," *Economic Record*, September 1955, *31*, 242–60.

______, *Trade Policy and Economic Welfare*, Oxford: Oxford University Press, 1974.

Corwin, A. F., *Immigrants — and Immigrants: Perspectives on Mexican Labor Migration to the United States*, Westport: Greenwood Press, 1978.

Dixit, Avinash and Norman, Victor, *Theory of International Trade*, London: Cambridge University Press, 1980.

Doeringer, Peter and Piore, Michael J., *Internal Labor Markets and Manpower Analysis*, Lexington: D.C. Heath, 1971.

Ethier, Wilfred J., "Dumping," *Journal of Political Economy*, June 1982, *90*, 487–506.

Gordon, David, Edwards, Richard and Reich, Michael, *Segmented Work, Divided Workers*, London: Cambridge University Press, 1982.

Hiemenz, U. and Schatz, K. W., *Trade in Place of Migration: An Employment-Oriented Study with Special Reference to the Federal Republic of Germany, Spain and Turkey*, Geneva: International Labor Organization, 1979.

Johnson, Harry G., *International Trade and Economic Growth*, Cambridge: Harvard University Press, 1965.

Kemp, Murray C., *The Pure Theory of International Trade and Investment*, Englewood Cliffs: Prentice-Hall, 1969.

Kennedy-Brenner, C., *Foreign Workers and Immigration Policy — The Case of France*, Paris: Organization for Economic Cooperation and Development, 1979.

Markusen, James R. and Melvin, James R., "Tariffs, Capital Mobility, and Foreign Ownership," *Journal of International Economics*, August 1979, *9*, 395–409.

Marshall, Ariadne, *The Import of Labor*, Rotterdam: Rotterdam University Press, 1973.

Ohlin, Bertil, *Interregional and International Trade*, rev. ed., Cambridge: Harvard University Press, 1967.

Piore, Michael J., *Birds of Passage: Migrant Labor and Industrial Societies*, London: Cambridge University Press, 1979.

Zachariah, K. C. and Conde, J., *Migration in West Africa: Demographic Aspects*, New York: Oxford University Press, 1981.

Can Small Deviations from Rationality Make Significant Differences to Economic Equilibria?

By GEORGE A. AKERLOF AND JANET L. YELLEN*

This paper concerns the robustness of economic equilibria in familiar economic models. It addresses the question whether a small amount of nonmaximizing behavior by agents is capable of causing changes in the equilibrium of the system which are an order of magnitude larger than the losses due to nonmaximization by the individual agents themselves.

For some, it would be reassuring to find that the results of models based on maximizing behavior continue to hold as an approximation when the assumption of strict maximization is relaxed just slightly. It is, after all, difficult to believe that agents *literally* maximize all the time. People may suffer from inertia. (Or, alternatively stated, they may have some small transactions costs in changing their actions.) They may also rely on rules of thumb which produce acceptable results on the average. So most economic theory based on strict maximization is useful when accompanied with the folk theorem that the results of this theory are approximately correct if the deviations from optimality (or the transactions cost of decision making) are small.

On the other hand, if slight relaxation of maximization results in rather different results from those of strict maximization, then there is the possibility of explaining phenomena that have been puzzling in the context of economic theory based on strict maximizing, such as the persistence of cartels and the existence of the business cycle.

This paper constructs a number of examples which show that small deviations from rationality *can* have first-order consequences in microeconomic analysis. Our 1985 paper shows that such deviations can account for Keynesian business cycles. All of these examples assume that a fraction β of the population fails to maximize. The rest of the population is totally rational. An equilibrium of such an economy is termed *near rational* if no nonmaximizer stands to gain a significant amount by becoming a maximizer. Technically, this means that the potential losses due to nonmaximization are second-order small (varying with the square of a shift parameter). Nonetheless we show that such behavior does commonly lead to changes in equilibrium in familiar models (relative to the equilibrium with full maximization) which are first-order, that is, that vary approximately proportionately with the shift parameter.

These results can be interpreted as a precise sense in which the conclusions usually derived from models with strict maximizing behavior are not very robust. We wish to emphasize that this theory does pass Lucas' test of a good model—that there are no opportunities for large increases in profits, or, in his terminology, that "there are no $500 bills on the sidewalk." To be specific, there is a fraction of the population who are strict maximizers. If there are opportunities for gain generally available (in Lucas' example, to pick up $500 off the sidewalk), a significant fraction of the population will

*University of California, Berkeley, CA 94720. We express our special gratitude to Robert Barro and Hajime Miyazaki for greatly elucidating comments on an earlier paper presented at the NBER-West Conference on Macroeconomics, December 2, 1983, and to John Haltiwanger and Michael Waldman for criticisms of an earlier draft. We also thank Andrew Abel, Kenneth Arrow, Sean Becketti, Ben Bernanke, Alan Blinder, Stephen Blough, Roger Craine, Avinash Dixit, Richard Gilbert, Paul Evans, Milton Friedman, Robert Hall, Bent Hansen, Robert Lucas, Thomas MaCurdy, John Roberts, Joseph Stiglitz, Steven Stoft, Lawrence Summers, James Tobin, and James Wilcox for valuable discussions which have led directly or indirectly to this paper. Our research was generously supported by National Science Foundation grant no. SES 81-19150, administered by the Institute for Business and Economic Research of the University of California-Berkeley.

readily avail itself of these opportunities. Additionally, the losses to those who are not maximizing are small, in a well-defined sense.

Recent results in industrial organization theory in the context of sequential games are similar in spirit. In recent interesting papers by Drew Fudenberg and Eric Maskin (1983), David Kreps et al. (1982), Kreps and Robert Wilson (1982), Paul Milgrom and John Roberts (1982), and Roy Radner (1980), it has been found that the presence of nonmaximizing behavior, even in rather small amounts, can significantly alter the equilibrium obtained. Other papers which examine the conditions under which nonmaximizing behavior can have a significant impact on the characteristics of economic equilibria include John Haltiwanger and Michael Waldman (1985) and Thomas Russell and Richard Thaler (1985). John Conlisk (1980) has examined conditions under which rule-of-thumb behavior will coexist with maximizing behavior in the long run.

Section I explains the basic logic of this paper, which can be generally expressed in terms of the envelope theorem. The next sections present four examples of equilibria that are near-rational, but that differ by a first-order amount from the equilibrium that would prevail with full rationality. Section II presents an example of a pure exchange economy. There is an initial (long-run) equilibrium of this economy in which all agents are exactly maximizing. Then each agent's endowment of one good is increased by a proportion ε. A fraction β of the population has inertial behavior in its consumption of that good while the rest of the population strictly maximizes. It is shown that this behavior by nonmaximizers results in individual losses which are second-order in ε, while there is (almost always) a first-order change (with respect to ε) in the distribution of income.

In the pure exchange economy, nonmaximization causes only a second-order movement from the Pareto frontier. But if there are externalities, nonmaximization which results in only second-order losses to the agents can cause first-order changes in the economy's deadweight loss. This proposition is illustrated by two examples in Section III.

Section IV presents an example of cartel formation. It is supposed initially that there is a Cournot equilibrium of an oligopolistic market. There is a posited reduction in output by a fraction ε by a proportion β of the firms, while the rest of the firms maximize (according to the Cournot assumption that the output produced by rival firms will remain unchanged.) With some qualification, it is shown that such a reduction increases the oligopolists' profits by a first-order amount while the expected gains from cheating on the cartel are second-order in ε.

I. Implications of the Envelope Theorem

There is a simple reason why deviations from rationality, that involve only second-order losses for nonmaximizing agents, nevertheless can cause first-order changes in the economy's equilibrium. The reasoning is based on the *envelope theorem* and its general equilibrium implications. The envelope theorem states, in effect, that first-order "errors" made by an agent in setting any decision variable, such as quantities produced or consumed, normally cause only second-order losses in utility or profit. If a significant fraction of the population makes such errors and their mean value is nonzero, the errors are likely to have general equilibrium effects which are first-order.

More formally,[1] consider the unconstrained maximization problem: maximize $f(x, a)$ where x is a choice variable and a is a vector of parameters or variables exogenous to the agent. Let $x(a)$ denote the unique maximizing choice of x, given a, and let $M(a) = f(x(a), a)$ denote the maximum value of f for given a. According to the envelope theorem:

$$(1) \quad dM(a)/da = \partial f(x(a), a)/\partial a.$$

In words, the envelope theorem states that, at the margin, the change in the objective function caused by a change in a is identical whether the agent adjusts x optimally or not at all. Inertial behavior is virtually costless.

[1] See Hal Varian (1978).

The theorem is easily extended for constrained maximization problems.[2]

The proof of the envelope theorem is trivial but instructive. Total differentiation of $M(a)$ yields

$$(2)\quad \frac{dM(a)}{da}=\frac{\partial f(x,a)}{\partial x}\frac{dx(a)}{da}+\partial f(x(a),a)/\partial a.$$

The first-order condition for an optimum requires that $\partial f(x(a),a)/\partial x=0$ and thus the envelope theorem follows. The first term in (2), that is the indirect effect of a change in a on f occurring via a change in optimal x, is zero. Thus the effect of a change in a on M is just the *direct* effect of a on f for given x.

A graphical interpretation of the theorem is given in Figure 1, which depicts f as a function of x for two alternative values of a, a_0, and a_1. The optimal choices of x are x_0 and x_1, respectively, and the maximized values of the objective function f are $M(a_0)=f(x_0,a_0)$ and $M(a_1)=f(x_1,a_1)$. An agent who behaves inertially following a change in a, leaving x fixed at x_0 instead of changing x to x_1 will incur a loss, $\mathscr{L}$, from his failure to maximize equal to $f(x_1,a_1)-f(x_0,a_1)$ which is obviously second-order small.

There is an intuitive explanation of this result. The essential feature of optimal choice is indifference at the margin. Take the case of a consumer purchasing apples and bananas. At an optimum, the last dollar spent on apples adds as much utility as the last dollar spent on bananas. If the consumer is forced to spend a small amount less than the optimum on apples, the small amount extra he or she can spend on bananas provides almost perfect compensation.

More formally, writing the loss as a Taylor-series expansion around x_1 yields

$$(3)\quad \mathscr{L}\simeq\tfrac{1}{2}\left(\partial^2 f(x_1,a_1)/\partial x^2\right)\big|_{x_1,a_1}(x_0-x_1)^2.$$

The loss from failure to maximize is proportional not to the error made in setting x, but to the square of the error. The loss can also be expressed in terms of the change in a, which can be thought of as a shock to the system. Let $\varepsilon=a_1-a_0$ denote the shock. Then

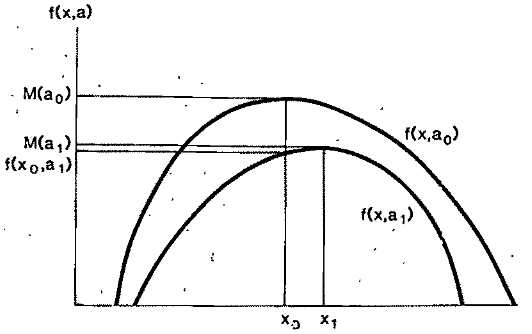


FIGURE 1

$$(4)\qquad x_1-x_0\simeq(dx(a)/da)\varepsilon$$

where $dx(a)/da$ normally differs from zero. Thus,[3]

$$(5)\qquad \mathscr{L}(\varepsilon)\simeq\mathscr{L}''(0)\varepsilon^2$$

where

$$\mathscr{L}''(0)=\frac{1}{2}\left[\frac{\partial^2 f(x,a)}{\partial x\,\partial a}\right]^2\bigg/\frac{\partial^2 f(x,a)}{\partial x^2}.$$

If there is a shock of size ε, an agent who behaves inertially makes an error which is ordinarily proportional to ε, or first-order, and incurs a loss which is proportion to ε^2, or second-order. Behavior that leads to a loss proportional to ε^2, where ε is a shock parameter, is called *near-rational.*

Although the envelope theorem is itself trivial, its implications are not. In many general equilibrium models, first-order errors in choice variables (for example quantities demanded or supplied) by a fraction of agents will cause first-order changes in such endogenous variables as relative prices. These endogenous variables typically enter the objec-

[2]See Varian (pp. 268–69.)

[3]This follows from the fact that $dx(a)/da=-(\partial^2 f/\partial x\,\partial a)/(\partial^2 f/\partial x^2)$.

tive functions of households and firms; they are arguments in the a vector and any change in these variables will accordingly have welfare consequences. As long as the errors do not just cancel out when averaged over the population (a sensible assumption in many contexts), "small" departures from rationality can cause first-order changes in such variables as prices, income distribution, or deadweight loss.

To be more precise, let p denote an endogenous variable (such as relative prices) which enters the a vector in a particular system under consideration. Assume that a fraction β of the population fail to maximize in a well-specified fashion. It is possible to compute the equilibrium value of p corresponding to alternative values of ε and β, so that p can be written as $p = p(\varepsilon, \beta)$. Let $s(\varepsilon, \beta)$ denote the difference between the equilibrium value of p with a fraction β of nonmaximizers and the equilibrium value of p with no nonmaximizers; that is, $s(\varepsilon, \beta) = p(\varepsilon, \beta) - p(\varepsilon, 0)$. The function $s(\varepsilon, \beta)$ gives the *systemic effect* of nonmaximizing behavior. Writing $s(\varepsilon, \beta)$ as a Taylor-series approximation around the point $(0, \beta)$ yields

$$(6) \quad s(\varepsilon, \beta) \simeq \left.\frac{\partial s(\varepsilon, \beta)}{\partial \varepsilon}\right|_{\varepsilon=0,\beta} \varepsilon + \frac{1}{2} \left.\frac{\partial^2 s(\varepsilon, \beta)}{\partial \varepsilon^2}\right|_{\varepsilon=0,\beta} \varepsilon^2$$

As we demonstrate by example in Sections II–IV, there are many models in which the individual losses due to the assumed deviations from maximization are approximately proportional to ε^2 (second-order), but the function $s(\varepsilon, \beta)$ has a first partial derivative that is nonvanishing. In such cases, the systemic effect of nonmaximization is approximately proportional to ε, or first-order; the systemic effect of nonmaximization is an order of magnitude larger than the individual losses. This implies that the ratio, denoted $R(\varepsilon, \beta)$, of the "systemic effect" of nonmaximizing behavior to the individual loss approaches infinity as ε approaches zero.

Methodology of the Examples. The methodology employed in each of the next sections is identical and can be outlined briefly. First, a simple model is constructed and its equilibrium is calculated under the assumption of full maximization by all agents. Second, the system is shocked in a specific way in amount ε. Third, the values of relevant endogenous variables are computed in the equilibrium of this model on the assumption that a fraction β of agents fail to maximize in a well-specified way. Finally, two questions are posed: first, is the assumed behavior by nonmaximizers near-rational? In all cases, the answer is yes and can be demonstrated by showing that the loss from failure to maximize is second-order: $\mathscr{L}(\varepsilon) \simeq \mathscr{L}''(0)\varepsilon^2$. Since it is always the case in these examples that $\mathscr{L}(0) = 0$, it is only necessary to demonstrate that $\mathscr{L}'(0) = 0$. This follows from the envelope theorem and in most cases proof will therefore be omitted. Second, does the equilibrium of the system corresponding to strict maximization by all agents ($\beta = 0$) differ by a first-order amount from the equilibrium of the system when there are a fraction β of nonmaximizers? This involves showing that $\partial s(\varepsilon, \beta)/\partial \varepsilon|_{\varepsilon=0} \neq 0$.

Some Caveats. Some caveats are in order concerning the appropriate interpretation of our results. In each example we show that the ratio of the systemic effect of nonmaximization to the individual loss from nonmaximization approaches infinity as ε approaches zero. The potential for applying the logic of the envelope theorem in any context where maximizing behavior normally occurs suggests that our theorem could have almost unlimited applications. But for the theorem to have practical relevance, it must be true for finite values of ε, corresponding to economically noticeable shocks, and not just for infinitesimal ε, that the ratio $R(\varepsilon, \beta)$ is "large." The behavior of R as ε increases depends critically on the magnitudes of $\mathscr{L}''(\varepsilon)$ and $\partial s(\varepsilon, \beta)/\partial \varepsilon$ in the neighborhood of $\varepsilon = 0$. If $\mathscr{L}''(0)$ is extremely large, or if the first partial of s is very small, the ratio of the systemic effect to the individual loss will diminish rapidly as ε increases, limiting the applicability of our infinitesimal results. The

simulations presented in Sections II and III provide a simple way of evaluating the behavior of this ratio for finite ε in some specific models for reasonable parameter choices. In these models the ratio of effects remains large for "significant" ε. Economic intuition suggests, however, that there will be circumstances in which $\mathscr{L}''(\varepsilon)$ is quite large or $\partial s(\varepsilon,\beta)/\partial\varepsilon$ small.[4]

Markets which are close to perfectly competitive provide an obvious example in which $\mathscr{L}''(\varepsilon)$ is large. Consider the case of a monopolistically competitive firm setting its price. If the firm charges a price above the optimum, sales and profits fall. In the limiting case of perfect competition, sales plummet to zero and profits drop discontinuously. A price below the optimum raises sales but lowers profits. In the limiting case of perfect competition in which potential sales at the market price are unlimited, price cuts lower profits by an amount which is proportional to the error in pricing. In effect, the function relating lost profits to price errors has an extremely large second derivative.

It is equally easy to envision cases in which nonmaximizing behavior by a significant fraction of the population would have only a minor effect on the equilibrium of the system so that $\partial s(\varepsilon,\beta)/\partial\varepsilon$ is small. Suppose, for example, that a significant minority of rational market participants regards two goods or assets as perfect substitutes at a given price ratio; the opinions of such agents is apt to dictate the market outcome even in the presence of a substantial fraction of irrational or nonmaximizing agents. Financial markets provide the obvious example of a context in which "efficient market" outcomes can be achieved as long as some rational arbitrageurs operate in the market. More generally, the recent work of Haltiwanger and Waldman shows that in any market which can be characterized as one with "congestion effects," it will be true that sophisticated or maximizing agents will have a disproportionately large effect on the equilibrium. In these circumstances, $\partial s(\varepsilon,\beta)/\partial\varepsilon$ is small. With the preceding words of warning, we now proceed to our first example.

[4]In our examples, for given parameter values there always exists a band of ε values such that $|R|$ exceeds any specified value within the band. However, as the referees have pointed out to us, in most examples, it is possible to select extreme parameter values which make $|R|$ arbitrarily small for any (ε,β) pair, $\varepsilon>0$.

II. Near-Rational Behavior in a Pure Exchange Economy

The first example concerns the canonical case of a pure exchange economy. A fraction of individuals are assumed to demand one good inertially, following an endowment shock. This form of nonmaximizing behavior is near-rational. The consequences of inertial behavior for income distribution are first-order, however. In comparison with the equilibrium in which all agents maximize, it turns out that some people are better off and some worse off. There is a first-order movement along the utility possibility frontier (but only a second-order movement away from it).

The Initial Equilibrium. Consider an economy with equal numbers of consumers of two types and two goods denoted by subscripts 1 and 2. Consumers of groups 1 and 2 have utility $U(x_1,x_2)$ and $V(y_1,y_2)$, respectively, and have initial endowments $(\bar{x}_1,\bar{x}_2)$ and $(\bar{y}_1,\bar{y}_2)$, respectively. Let p denote the relative price of good 2. The equilibrium value of p is determined by the condition that excess demand for each good be zero. Let the superscript o denote the values of variables in the initial (long-run) equilibrium, in which all agents maximize utility.

The Shock. It is then assumed that an endowment shock occurs. Many different alternatives could be considered. We consider the special case in which each individual's endowment of good 2 increases by a fraction ε. The new endowment vectors are accordingly $(\bar{x}_1,\bar{x}_2(1+\varepsilon))$, $(\bar{y}_1,\bar{y}_2(1+\varepsilon))$. A fraction β of both type-1 and type-2 consumers have inertial consumption of good 2, their consumption being x_2^o and y_2^o, respectively.

Two propositions will be demonstrated. According to Proposition 1, the individual

loss due to nonmaximization in this example is second-order. This can be stated more formally.

PROPOSITION 1:

$$(7)\qquad dU^m/d\varepsilon|_{\varepsilon=0} = dU^n/d\varepsilon|_{\varepsilon=0}.$$

Similarly,

$$(8)\qquad dV^m/d\varepsilon|_{\varepsilon=0} = dV^n/d\varepsilon|_{\varepsilon=0},$$

where $U^m(\varepsilon;\beta)$ and $V^m(\varepsilon;\beta)$ denote values of utility in equilibrium for maximizers of type 1 and type 2 as a function of ε for given β; and where $U^n(\varepsilon;\beta)$ and $V^n(\varepsilon;\beta)$ are the values of utility in equilibrium of nonmaximizers of type 1 and type 2, respectively, as a function of ε for given β.

PROOF:

Let $\mathscr{U}(p, I_1)$ denote the indirect utility function of a type 1 individual. This function gives the maximum attainable utility corresponding to alternative prices, p, and his income, I_1, measured in terms of good 1. We can thus write:

$$(9)\quad U^m(\varepsilon;\beta) = \mathscr{U}(p(\varepsilon;\beta), \bar{x}_1 + p(\varepsilon;\beta)\bar{x}_2(1+\varepsilon)).$$

where $p(\varepsilon;\beta)$ is the equilibrium value of price as a function of ε for given β. Differentiating with respect to ε and applying Roy's Identity yields

$$(10)\quad dU^m/d\varepsilon|_{\varepsilon=0} = \lambda_1[p^o\bar{x}_2 + (\bar{x}_2 - x_2^o)(dp/d\varepsilon)|_{\varepsilon=0}],$$

where λ_1 is the marginal utility of income of a type-1 consumer. Equation (10) shows that the welfare effect of a small endowment shock consists of two parts: the first part, $p^o\bar{x}_2$ is the value of the extra endowment evaluated at the initial price p^o. The second part, $(\bar{x}_2 - x_2^o)(dp/d\varepsilon)|_{\varepsilon=0}$, is the income effect of the price change caused by the endowment shock. Assuming that $dp/d\varepsilon$ is negative, net sellers of good 2 (for whom $(\bar{x}_2 - x_2^o)$ is greater than zero) lose from the price change; net purchasers gain.

How does $dU^m/d\varepsilon$ as given by (10) compare to the welfare effect of the endowment shock on a type-1 nonmaximizer, who demands good 2 inertially? He consumes $x_2 = x_2^o$ and spends the rest of his income on good 1 so that his utility is

$$(11)\quad U^n(\varepsilon;\beta) = U(\bar{x}_1 + p(\varepsilon;\beta)\cdot(\bar{x}_2(1+\varepsilon) - x_2^o), x_2^o).$$

Differentiating with respect to ε at $\varepsilon = 0$, yields[5]

$$(12)\quad (dU^n/d\varepsilon)|_{\varepsilon=0} = (dU^m/d\varepsilon)|_{\varepsilon=0} = \lambda_1[p^o\bar{x}_2 + (dp/d\varepsilon)|_{\varepsilon=0}(\bar{x}_2 - x_2)].$$

It remains to show that inertial behavior by a fraction β of the population has a systemic effect, in this case on the equilibrium prices in the economy and the welfare of all agents, which is first-order in ε.

Inspection of (10) indicates that nonmaximizing behavior affects welfare only by altering the price effect, $dp(\varepsilon;\beta)/d\varepsilon|_{\varepsilon=0}$. If there is trade in the initial equilibrium (so that $\bar{x}_2 - x_2^o = y_2^o - \bar{y}_2 \neq 0$) nonmaximization has first-order real income effects provided that Proposition 2 is valid:

PROPOSITION 2:

$$(dp(\varepsilon;\beta)/d\varepsilon - dp(\varepsilon;0)/d\varepsilon)|_{\varepsilon=0} \neq 0.$$

PROOF:

The equilibrium price, $p(\varepsilon;\beta)$, is determined by the condition that the total demand for good 2 equals the total supply of good 2:

$$(13)\quad \beta(x_2^o + y_2^o) + (1-\beta)x_2(p, I_1) + (1-\beta)y_2(p, I_2) = (\bar{x}_2 + \bar{y}_2)(1+\varepsilon),$$

where I_1, I_2 measure the income of type-1

[5] $p(0,\beta) = p^o$ for all β.

and type-2 consumers in terms of good 1, and where $x_2(p, I_1)$, $y_2(p, I_2)$ denote the demand for good 2 by maximizers of types-1 and 2, respectively.

Differentiation of (13) yields:

$$(14)\quad \left.\frac{dp}{d\varepsilon}\right|_{\varepsilon=0} = \frac{G+(\bar{x}_2+\bar{y}_2)\beta/(1-\beta)}{H}$$

where $G = \bar{x}_2\left(1-p^o\frac{\partial x_2}{\partial I_1}\right)+\bar{y}_2\left(1-p^o\frac{\partial y_2}{\partial I_2}\right)$

and $H = \frac{\partial x_2}{\partial p}+\frac{\partial y_2}{\partial p}+\frac{\partial x_2}{\partial I_1}\bar{x}_2+\frac{\partial y_2}{\partial I_2}\bar{y}_2.$

In what we shall refer to as the "normal" case, the marginal propensity to consume each good is less than one, and the excess demand for good 2 is inversely related to its price. As a result of these two properties, in the normal case $G>0$ and $H<0$.

From (14) the effect of nonmaximization on price can be calculated as:

$$(15)\quad \left.\frac{dp(\varepsilon;\beta)}{d\varepsilon}\right|_{\varepsilon=0} - \left.\frac{dp(\varepsilon;0)}{d\varepsilon}\right|_{\varepsilon=0} = \beta(\bar{x}_2+\bar{y}_2)/(1-\beta)H.$$

This last expression (15) is not equal to zero. The quantity G is the initial excess supply induced by the endowment shock when all agents maximize. The quantity H is the change in excess demand for good 2 induced by a change in price when all agents maximize. When a fraction β of agents have inertial consumption of good 2, as in this model, it is as though the initial excess supply were larger than G by an amount $\beta(\bar{x}_2+\bar{y}_2)/(1-\beta)$. In the normal case, nonmaximization causes a larger reduction in the price of good 2 with corresponding gains for net purchasers of good 2 and losses for net suppliers of good 2.[6] These welfare effects are first-order moves along the economy's utility possibility frontier, but are not first-order departures from the frontier.

Table 1 gives the results of simulations where Group 1 have utility function $U = x_1^{\alpha}x_2^{1-\alpha}$ and Group 2 have utility function $V = y_1^{\gamma}y_2^{1-\gamma}$. In this example initial endowments are chosen so that there is considerable trade between the two groups and therefore the income effects of price changes are nonnegligible. Accordingly Group 1 are endowed only with good 2 and Group 2 are endowed only with good 1, so that initially, Group 1 are net buyers of good 1.

The table shows the consequences of a 1 percent increase in the endowment of good 2 for various values of α and γ. The "*change* in group utility" gives the difference in percentage terms between the average utility of group members when a fraction β of the population are nonmaximizers and when $\beta = 0$. Since Group 1 individuals are sellers of good 2, they lose as β increases because the fall in p is greater the larger is β. Inspection of the table shows that the ratio between the change in group utility and the "*loss* to nonmaximizers" is normally quite large.

III. The Welfare Effects of Near-Rational Behavior in the Presence of Distortions

Although near-rational behavior can cause first-order changes in the distribution of income or utility in an economy without distortions, it cannot cause departures from Pareto optimality. Two examples presented in this section show that near-rational behavior *can* cause first-order changes in social welfare in an economy with distortions (for example, externalities, monopoly, taxes). The first example demonstrates that a near-rational, as opposed to a strictly rational, response to technical change by a monopolist, can cause a first-order change in deadweight loss. The second example concerns a monetary economy. In such an economy, with near-rational inertial demand for money by a fraction of the population, a money rain results in a first-order change in aggregate welfare. The externality occurs in this model because the utility of an agent with a given final consumption bundle of goods and money is dependent on the price

[6] There is an interesting special case where inertial behavior causes no change in the equilibrium. This occurs in the example of this section if consumption of good 1, rather than good 2, is adjusted inertially.

TABLE 1—PERCENTAGE LOSS IN AVERAGE UTILITY AND PERCENTAGE LOSS IN UTILITY DUE TO NONMAXIMIZING BEHAVIOR FOLLOWING A 1 PERCENT INCREASE IN THE ENDOWMENT OF GOOD 2 ($\varepsilon = 1.0$ PERCENT) FOR DIFFERENT GROUP EXPENDITURE SHARES (α, γ) AND FRACTIONS OF NONMAXIMIZERS (β)[a]

	$\beta = .25$				$\beta = .50$				$\beta = .75$			
	Group 1		Group 2		Group 1		Group 2		Group 1		Group 2	
	Change[b]	*Loss*[c]	*Change*	*Loss*	*Change*	*Loss*	*Change*	*Loss*	*Change*	*Loss*	*Change*	*Loss*
$\alpha = .25$												
$\gamma = .25$	−.331	.015	.969	.076	−.973	015	2.791	.321	−2.778	.015	7.366	1.769
$\gamma = .50$	−.331	.015	.651	.026	−.973	.015	1.903	.114	−2.778	.015	5.247	.668
$\gamma = .75$	−.331	.015	.326	.009	−.973	.015	.956	.039	−2.778	.015	2.667	.233
$\alpha = .50$												
$\gamma = .25$	−.330	.005	.485	.039	−.978	.005	1.419	.123	−2.848	.005	3.959	.596
$\gamma = .50$	−.330	.005	.326	.013	−.978	.005	.964	.042	−2.848	.005	2.762	.215
$\gamma = .75$	−.330	.005	.163	.004	−.978	.005	.484	.014	−2.848	.005	1.397	.073
$\alpha = .75$												
$\gamma = .25$	−.329	.002	.322	.030	−.980	.002	.950	.076	−2.872	.002	2.708	.321
$\gamma = .50$	−.329	.002	.217	.010	−.980	.002	.645	.026	−2.872	.002	1.874	.114
$\gamma = .75$	−.329	.002	.109	.003	−.980	.002	.324	.009	−2.872	.002	.946	.039

[a]Endowments: $\bar{x}_1 = 0$, $\bar{x}_2 = 100$; $\bar{y}_1 = 100$, $\bar{y}_2 = 0$; Utility Functions: both groups are assumed to possess Cobb-Douglas utility functions with good 1 expenditure shares of α and γ, respectively.

[b]Change in Group Utility (*Change*) is the percentage difference between the average utility of group i individuals for the given β and the utility of the same group when $\beta = 0$.

[c]Loss of Nonmaximizers (*Loss*) is the percentage by which a nonmaximizer could increase his utility by selecting the consumption bundle optimally.

level. In the standard model without money, an agent's utility is only dependent upon his final consumption bundle.

A. *Monopoly and Technical Change*

The Initial Equilibrium. Figure 2 illustrates the consequences of technical change in a monopolistic industry. The monopolist has a linear demand curve $p = a - bQ$ and produces output with constant marginal cost equal to average cost, initially equal to c. Initially the monopolist sets price to maximize profit (where marginal revenue equals marginal cost.)

The Shock. Now assume that a technical change occurs which reduces the monopolist's costs by a fraction ε to $c' = c(1 - \varepsilon)$. If the monopolist maximizes following the shock, his output will expand from Q^o to Q^1. The change in output $Q^1 - Q^o = c\varepsilon/2b$ is proportional to the shock. The new deadweight loss is given by the triangle *CHE* in Figure 2.

Suppose, however, that the monopolist had kept price and output unchanged at (p^o, Q^o) following the shock. The monopolist's loss in

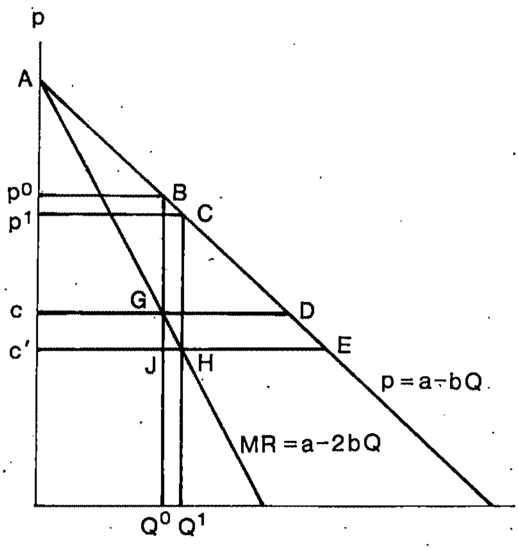


FIGURE 2

profit from failure to maximize is given by the area of the small triangle *GJH*, which is $c^2\varepsilon^2/4b$. This loss is second-order because each extra unit of output produced by the maximizing monopolist raises profitability by the difference between marginal revenue and marginal cost, which were initially equal.

In contrast, the social loss from the failure by the nonmaximizing monopolist to expand output is an order of magnitude larger than the private loss. Each extra unit of output sold creates social surplus equal to the gap between the price and marginal cost. The expansion of output by the maximizing monopolist by $c\varepsilon/2b$ units following the technical change results in social surplus in amount $BJHC$, which is approximately equal to the product of the difference between the initial price and cost and the expansion of output, or $(p^o - c)(c\varepsilon/2b)$. Geometrically, $BJHC$ is the difference between the total deadweight loss in the nonmaximizing case (the triangle BJE) and the total deadweight loss in the maximizing case (the triangle CHE).

To summarize, the behavior by the nonmaximizing monopolist is near-rational, but the social loss from that behavior is first-order.

B. *The Welfare Consequences of Money Supply Shocks*

The next example of near-rational behavior is motivated by a finding that is consistently obtained in empirical research on portfolio choice. Economists who have estimated portfolio balance models, relating asset holdings to the theoretically predicted determinants of optimal portfolio shares such as expected rates of return, the price level, and real income, routinely find long lags in the adjustment of actual to optimal holdings. For example, Stephen Goldfeld (1976) finds that the short-run partial elasticity of money demand with respect to the price level is .27, while theory offers the prediction that the elasticity should be unity. Sluggish adjustment of asset holdings seems perplexing since financial markets are usually thought to have low transaction costs. A possible explanation for inertial behavior is that the losses from failure to optimize are so small that people are slow to make the adjustments required by full optimization, permitting portfolio proportions to diverge from the optimizing levels for significant periods. If individuals are indeed near-rational in their portfolio behavior, the logic of this paper suggests that such behavior can have systemic repercussions which affect social welfare. It is possible to demonstrate the validity of this conjecture by means of an example based on a rudimentary model in the style of Patinkin.

The Initial Equilibrium. Consider a pure exchange economy with identical agents. Each agent has an initial endowment of a single commodity called "goods" in amount $\overline{G}$ and money in amount $\overline{M}_o$. All agents have the same utility function:

$$U = (M/p)^{\alpha} G^{1-\alpha}, \tag{16}$$

where p is the price at which money is traded for goods, M is money held, and G is goods consumed. Assume that initially the economy is in a "long-run equilibrium" with all agents choosing M and G to maximize utility subject to the budget constraint:

$$M + pG = \overline{M}_o + p\overline{G}. \tag{17}$$

The condition that the goods and money markets clear permits the characterization of the long-run equilibrium. The initial price level p^o is

$$p^o = (1-\alpha)\overline{M}_o/\alpha\overline{G}. \tag{18}$$

The Shock. Now consider the consequences of a "money rain" resulting in an increase in the money supply by a fraction ε with each agent receiving a proportional share. Such an operation is completely neutral if all agents optimize. The price level merely increases by a fraction ε, real cash balances remain unchanged, and the utility of all agents remains unchanged. The failure of a fraction β of the population to choose money holdings optimally changes the analysis, causing a money rain to have a negative effect on real welfare. Specifically, assume that a fraction β of the population demand money inertially in the short run, keeping money holdings constant at $\overline{M}_o$ in the face of changes in the price level.[7] The remaining fraction of the population choose money balances, given the price

[7]Akerlof (1979) has shown that with constant target-threshold monitoring, money demand will be independent of the price level.

TABLE 2—PERCENTAGE LOSS IN AVERAGE UTILITY AND PERCENTAGE LOSS IN UTILITY DUE TO NONMAXIMIZING BEHAVIOR FOR DIFFERENT CHANGES IN PRICES (θ); ELASTICITY OF UTILITY WITH RESPECT TO REAL BALANCES (α); AND FRACTION OF NONMAXIMIZERS (β)

	$\theta=.05$			$\theta=.10$		
	Percent Loss in Average Utility[a] (1)	Percent Loss Due to Nonmaximizing Behavior[b] (2)	Ratio of (1) to (2)	Percent Loss in Average Utility[a] (1)	Percent Loss Due to Nonmaximizing Behavior[b] (2)	Ratio of (1) to (2)
$\alpha=.1$						
$\beta=.25$	.132	.012	11.0	.257	.046	5.6
$\beta=.50$	.256	.012	21.3	.500	.044	11.4
$\beta=.75$	.375	.011	34.1	.730	.042	17.4
$\alpha=.3$						
$\beta=.25$	.471	.041	11.5	.917	.152	6.0
$\beta=.50$	.857	.034	25.2	1.667	.127	13.1
$\beta=.75$	1.180	.029	40.7	2.291	.109	21.0
$\alpha=.5$						
$\beta=.25$	.971	.074	13.1	1.886	.275	6.9
$\beta=.50$	1.613	.052	31.0	3.125	.196	15.9
$\beta=.75$	2.069	.039	53.1	4.000	.146	27.4

[a]Percentage loss in average utility equals $((\overline{U}-U_o)/U_o)\times 100$.
[b]Percentage loss in utility due to nonmaximizing behavior equals $((U^m-U^n)/U^n)\times 100$.

level p, to maximize utility. Optimal money holdings are

$$M^m=\alpha(p\overline{G}+\overline{M}_o(1+\varepsilon)). \tag{19}$$

The equilibrium condition that the demand for money by maximizers and nonmaximizers must equal supply, following the shock, determines the new equilibrium price level, $p(\varepsilon,\beta)$:

$$\beta\overline{M}_o+(1-\beta)\alpha[p(\varepsilon,\beta)\overline{G}+\overline{M}_o(1+\varepsilon)]=\overline{M}_o(1+\varepsilon). \tag{20}$$

Solving for the change in the price level, θ, as a function of ε yields

$$\theta=(p(\varepsilon,\beta)-p^o)/p^o = \frac{1-\alpha(1-\beta)}{(1-\beta)-\alpha(1-\beta)}\varepsilon>\varepsilon \quad \text{for } 0<\beta<1. \tag{21}$$

The failure of some agents to add to their money balances means that the price level must change by enough to induce the remaining fraction to hold all the extra money. As a result, the price level rises by a larger percentage than the money supply increases. Real balances per capita are lower in the new equilibrium. The percentage decline in average real cash balances is $\varepsilon(d\theta/d\varepsilon-1)$.

Defining the loss from nonmaximization as $\mathscr{L}(\varepsilon)=U^m-U^n$, it follows from the envelope theorem, or can be verified by tedious computation, that

$$d\mathscr{L}/d\varepsilon|_{\varepsilon=0}=0. \tag{22}$$

The individual loss due to nonmaximization is second-order. In contrast, the reduction in the equilibrium level of per capita real cash balances due to nonmaximization causes a first-order loss in average utility, $\overline{U}$, defined as a weighted average of the utilities of maximizers and nonmaximizers. Because the difference between U^n and U^m is second-order, average utility is, to a first-order of approximation, equal to the utility of a hypothetical person with average money holdings and the average holdings of goods. In the new equilibrium, average goods consumed remain unchanged at $\overline{G}$, but average real balances decline by $\varepsilon(d\theta/d\varepsilon-1)$. As a result, social loss, which is the percentage change in average utility due to a 1 percent

increase in the money supply, is approximated by:

$$(23)\quad -\alpha\left(\frac{d\theta}{d\varepsilon}-1\right)=-\alpha\beta/(1-\alpha)(1-\beta);$$

the last expression is the product of the elasticity of utility with respect to money balances and the percentage change in money balances due to a 1 percent change in the money supply.

Simulations. For various values of α and β, Table 2 gives the percentage loss in average utility when ε is selected to produce a price level change of 5 and 10 percent, respectively; this is compared to the percentage loss in utility suffered by individual nonmaximizers as a consequence of their failure to optimize. These individual losses are everywhere extremely small while the ratio of the two losses is quite large.

IV. A Model of Partial Cartel Formation

In the absence of sanctions for cheating, it is generally agreed that firms in an oligopolistic industry producing a homogeneous output over any finite time horizon are unlikely to be able to reduce output to the jointly profit-maximizing level. If firms try to collude to restrict output and earn significantly higher profit, it is argued that the large gap between marginal cost and marginal revenue at that point will create a strong incentive to cheat. As is well known, with perfect information and strictly maximizing behavior by all firms, the cartel outcome cannot be sustained as a noncooperative Nash-Cournot equilibrium if the industry's life is finite. This section shows, however, the existence of opportunities for cartel formation which restrict output enough to significantly raise the profits of cartel members which are nevertheless "viable" in the sense that the potential gain from cheating is second-order small (although not nonexistent).

The Initial Equilibrium. Assume that n identical firms produce a homogeneous output for a single period. Industry demand is $p=a-bQ$. Each firm has the same constant marginal = average cost = c. Each firm is assumed to follow the Cournot strategy of producing that output which is optimal given the outputs selected by rivals. In the unique Nash-Cournot equilibrium, the output of each firm q_i^o, profits Π_i^o and the price p^o are

$$(24)\quad q_i^o=(a-c)/b(1+n),$$

$$(25)\quad \Pi_i^o=(1/b)((a-c)/(1+n))^2,$$

$$(26)\quad p^o=(a+nc)/(1+n).$$

The Shock. Now suppose that a cartel is formed in the industry. In this model the shock is just a voluntary output cutback by firms in the cartel, as in the case of the 1973 oil shock. Members of the cartel cut back output by a fraction ε. A fraction β of the firms are noncheaters and producers the prescribed output, q_i^n:

$$(27)\quad q_i^n=(1-\varepsilon)(a-c)/b(1+n).$$

Assume that the remaining firms decide to cheat instead of adhering to the cartel agreement. They select their output to maximize profit in the usual Cournot fashion, choosing that output q_i^m, which maximizes profit given the outputs of all other maximizers and all nonmaximizers. The optimal output for a "maximizer" (cheater) satisfies the first-order condition that the marginal return to one additional unit of output, given the output of other firms is zero:

$$(28)\quad d\Pi_m/dq_i^m=a-bQ-bq_i^m-c=0,$$

where $Q=\beta nq_i^n+(1-\beta)nq_i^m$ = aggregate output. Solving for q_i^m yields

$$(29)\quad q_i^m=\frac{a-c-b\beta nq_i^n}{b[1+(1-\beta)n]}>q_i^o,$$

where q_i^n is given by (27). Each firm that cheats produces more than the Nash-Cournot equilibrium output, q_i^o, but aggregate industry output falls nevertheless. Straight-

forward calculation shows

$$(30)\quad \left.\frac{dQ}{d\varepsilon}\right|_{\varepsilon=0}=\frac{-\beta n(a-c)}{b(1+n)[1+n(1-\beta)]}<0,$$

$$(31)\quad \left.\frac{dp}{d\varepsilon}\right|_{\varepsilon=0}=\frac{(a-c)\beta n}{(1+n)[1+n(1-\beta)]}>0.$$

Profits of maximizers, who produce more and sell their output at a higher price, rise unambiguously. Nonmaximizers produce less but sell their output at a higher price. Their profits will also rise if they are sufficiently numerous. The exact condition is

$$(32)\quad \left.\frac{d\Pi_n}{d\varepsilon}\right|_{\varepsilon=0}\gtreqless 0 \quad \text{as} \quad \beta \gtreqless \frac{1+n}{2n}$$

For the noncheaters to gain they must be at least half of the industry.

These results show that the partial cartel can have a first-order effect on industry profits and potentially raises the profits of all firms by a first-order amount, even with a significant fraction of firms cheating.

Reward to a Nonmaximizer from Defecting. It remains to show that the reward to a nonmaximizer from defecting is second-order small. Let Π_n^* denote the "perceived" maximum profit which a typical nonmaximizer would potentially gain by altering his output to the optimum. Letting $Q_{n-1}=(1-\beta)nq_i^m+(\beta n-1)q_i^n$ denote the output of all firms in the industry with the exception of the potential cartel defector, the defector contemplates choosing a new output, q_n^*, to maximize his profit on the assumption that Q_{n-1} will remain unchanged. The optimal defection output satisfies a condition comparable to (28):

$$(33)\quad a-b\left(Q_{n-1}+q_n^*\right)-c-bq_n^*=0.$$

The potential defector knows that the new price resulting from his defection is

$$(34)\quad p^*=a-b\left(Q_{n-1}+q_n^*\right),$$

so that $\Pi_n^*=(p^*-c)q_n^*$. Defining the expected loss from not cheating as $\mathscr{L}(\varepsilon)=\Pi_n^*-\Pi_n$ it is straightforward to show that

$$(35)\quad \left.\frac{d\mathscr{L}}{d\varepsilon}\right|_{\varepsilon=0}=\left.\frac{d\left(\Pi_n^*-\Pi_n\right)}{d\varepsilon}\right|_{\varepsilon=0}=0.$$

The expected gain from defecting from the cartel is second-order small.

For a numerical illustration, let $a=10$, $c=0$, $b=1$, and $n=9$. In the Nash-Cournot equilibrium $q_i^o=\Pi_i^o=1$; $p^o=1$ and $Q=9$. Assume that a cartel is formed and members agree to cut back output by 10 percent so that $q_i^n=.9$. Assume that 7 firms are noncheaters, but 2 firms cheat. The cheaters each produce $q_i^m=1.233$. Cheaters earn profit of 1.52, gaining 52 percent. Noncheaters earn profit of 1.11 for a gain of 11 percent. A potential defector from the cartel would find it optimal to produce $q_n^*=1.066$. His defection would lower the price from 1.233 to 1.066. His potential profit, Π_n^*, is thus 1.138, which represents a gain of only 2.5 percent in profit.

V. Conclusion

This paper has presented a method for analyzing some implications of near-rational nonmaximizing behavior. According to the basic result, in a disturbance of an equilibrium in which all agents fully maximize, if many agents engage in the same type of nonmaximizing behavior, their individual losses, due to the envelope theorem, will be second-order (in terms of the parameter describing the disturbance) while the effect of their nonmaximizing behavior on the resultant equilibrium is frequently first-order. The four examples above illustrate the application of this result to microeconomics.

REFERENCES

Akerlof, George A., "Irving Fisher on His Head: The Consequences of Constant Target-Threshold Monitoring of Money Holdings," *Quarterly Journal of Economics*, May 1979, *93*, 169–88.

______ **and Yellen, Janet**, "A Near-Rational Model of the Business Cycle, with Wage and Price Inertia," *Quarterly Journal of*

Economics, forthcoming 1985.

Conlisk, John, "Costly Optimizers versus Cheap Imitators," *Journal of Economic Behavior and Organization*, 1980, *1*, 275–93.

Fudenberg, Drew and Maskin, Eric, "Folk Theorems for Repeated Games with Discounting and with Incomplete Information," mimeo., 1983.

Goldfeld, Stephen M., "The Case of the Missing Money," *Brookings Papers on Economic Activity*, *3*: 1976, 577–638.

Haltiwanger, John and Waldman, Michael, "Rational Expectations and the Limits of Rationality: An Analysis of Heterogeneity," *American Economic Review*, June 1985, *75*, 326–40.

Kreps, David M. and Wilson, Robert, "Reputation and Imperfect Information," *Journal of Economic Theory*, August 1982, *27*, 253–79.

Kreps, David M. et al., "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma," *Journal of Economic Theory*, August 1982, *27*, 245–52.

Milgrom, Paul and Roberts, John, "Predation, Reputation, and Entry Deterrence," *Journal of Economic Theory*, August 1982, *27*, 280–312.

Radner, Roy, "Collusive Behavior in Noncooperative Epsilon-Equilibria of Oligopolies with Long but Finite Lives," *Journal of Economic Theory*, April 1980, *22*, 136–54.

Russell, Thomas and Thaler, Richard, "The Relevance of Quasi Rationality in Competitive Markets," *American Economic Review*, December 1985, forthcoming.

Varian, Hal, *Microeconomic Analysis*, New York: Norton, 1978.

Employment Tenure and Earnings Profiles in Japan and the United States

By MASANORI HASHIMOTO AND JOHN RAISIAN*

A long-term employment relationship in Japan has been cited often as a principal reason for Japan's high labor productivity. Yet, recent discussions suggest that Japan may not be unique in enjoying strong employer-employee attachment. What are the facts? Are they consistent with the existing theories of long-term employment? These questions motivate the analysis of this paper. In light of the controversy about whether employment tenure differs between the two countries, an analysis is required as well for earnings profiles. Are they different? Are observed differences in tenure and in earnings profiles consistent with each other? Reexamination of the evidence leads us to conclude that, contrary to the impression created by recent literature, long-term employment is more prevalent in Japan than in the United States. Also, earnings-tenure profiles are more steeply sloped in Japan than in the United States. Finally, employment tenure in small Japanese firms is not universally short, as popular discussions suggest.

The paper is organized as follows. In Section I, we compare employment tenures in Japan and the United States. We also examine firm-size differences in employment tenure for the two countries. To assess the consistency of our findings with the recent theories on employment contracts, we examine differences in earnings profiles between Japan and the United States in Section II. The paper ends with a summary and concluding remarks in Section III.

I. Employer-Employee Attachments in Japan and the United States

The practice of lifetime employment in Japan has been viewed as a symbol of unique industrial relations in that country. Under this practice, an employee is hired by a firm immediately upon graduating from school, receives training on the job, and remains with the same employer until his retirement. Recently, journalists and academicians have taken issue with this.[1] Some have argued that the prevalence of lifetime employment in Japan is much too exaggerated. They point out that such practice applies to only about 30 percent of the male labor force. Furthermore, most workers are forced into retirement at age 55 with only modest company pensions and public social security that does not begin until the age 60.[2] Others have

*Indiana University, Bloomington, IN 47405, and Unicon Research Corporation, 10801 National Blvd, Los Angeles, CA 90064, respectively. The bulk of this research was conducted while Hashimoto was National Fellow at the Hoover Institution on leave from the University of Washington, and Raisian was Director of Research, Office of Policy, U.S. Department of Labor. Thanks are due Masahiko Aoki, Gary Becker, Barbara Brugman, Chris Hall, Keith Leffler, Thomas Moore, Junji Shiba, Hong Tan, Ben Yu, and an anonymous referee for useful comments. Research assistance at various stages by Barbara Brugman, Ed Fu, Jungi Shiba, Prapan Tianwattanatada, Charlotte Toney, and Bill Welch is acknowledged with thanks. Shoichi Ito of the University of Osaka Prefecture and Yuzuru Tsuzuki of the Japan Ministry of Labor were helpful in providing some data. Seminar participants at University of California-Los Angeles, U.S. Department of Labor, University of Chicago, University of North Carolina, and University of Rochester offered useful comments. We assume sole responsibility for any remaining errors. Hashimoto's research on comparative labor markets was financed in part by the U.S. Department of Labor contract no. J-9-M-2-0122 and the National Fellows Program at the Hoover Institution. Points of views stated in this paper do not necessarily represent the official position or policy of the U.S. Department of Labor or the Hoover Institution.

[1]See, for example, B. Bruce-Briggs (1982), Mitsuo Tajima (1982), Kazuo Koike (1977), Robert Hall (1982), and Walter Oi (1983a,b).

[2]See Bruce-Briggs. Note, however, that the mandatory retirement age in Japan has been rising steadily during the past ten years or so, though most retirement still occurs at ages below 60. In Japan, payment of lump sum amounts at the time of retirement rather than an

noted that long-term employment is much more prevalent in the United States than one might have thought, and even more prevalent than in Japan (for example, Robert Hall). A nonspecialist trying to ascertain the facts about lifetime employment in the two countries might find the recent literature confusing, but on balance would likely view the reported evidence as pointing to the conclusion that long-term employment is not unique to Japan.

The idea that long-term employment may be more prevalent in the United States than in Japan originates in Kazuo Koike's study (1977; see also 1978), which is cited by at least two prominent U.S. economists. Hall relies on Koike's evidence in commenting that "...tenure of fifteen years or longer is actually more common in the United States, in spite of the celebrated *nenko* system of lifetime employment in Japan" (p. 717). Walter Oi also refers to Koike's evidence when he says: "The myth of the protected and coddled Japanese worker was so pervasive that I never questioned its validity. A very different picture is painted by Koike..." (1983b, p. 70). It is natural to begin by reexamining Koike's evidence, which has generated much interest. To be fair, it should be noted that the evidence in question occupies only a small portion of Koike's extensive and valuable study.

Figure 1 reproduces Koike's data showing that the percent of male workers with fifteen or more years of tenure is indeed greater in the United States than in Japan. We think that this comparison is misleading because the Japanese data is for 1962. The permanent employment practice became widespread in Japan only after the mid-1950's. Before then long-term employment was limited to highly skilled and management-level workers in large firms. It is not surprising, then, to find that the proportion of workers with fifteen years or more tenure is not large in 1962. Postwar confusion in the Japanese economy makes the pattern observed by Koike even more plausible. A proper comparison must be based on more recent data for Japan than 1962. To this end, Figure 2 compares the distribution of tenure for the two countries in the late 1970's. The comparison of the two figures clearly indicates that long-term employment is distinctly more common in Japan than in the United States.[3] This conclusion is further corroborated by other evidence discussed below.

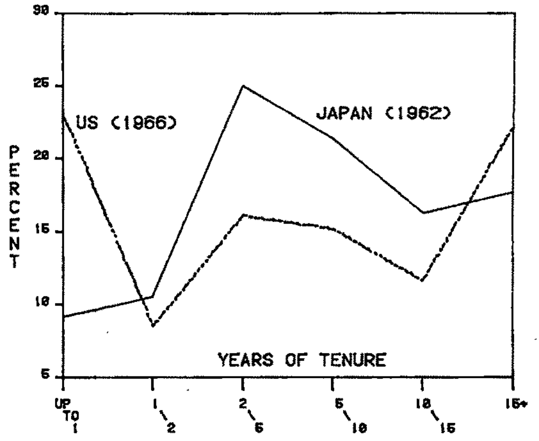


FIGURE 1

Table 1 compares the prevalence of long-term employment in the two countries. Column 6 lists the fifteen-year job retention rates. These are the estimated probabilities that males in a given age-tenure category will be with the same employer fifteen years later.[4] The rates are calculated as the proportion of workers in a given age-tenure category in the initial year with a correspondingly higher age and tenure fifteen years later. Although the Japanese employment surveys only infre-

annuity is common practice. Our preliminary examination reveals that the average lump sum payment is not much different from the present value of average pensions given to American workers. Since the Japanese retire earlier than Americans, however, annual payments are smaller in Japan.

[3] It is worth noting that even in earlier years (see Figure 1), the proportion of 40 to 54-year-old males with 15+ years of tenure is greater in Japan (43.4 percent) than in the United States (34.1 percent). In Figure 2, the tenure values of (2–5), (5–10), and (10–15) for Japan actually indicate (3–4), (5–9), and (10–14), respectively. This discrepancy is due to the fact that the published data are grouped differently for the two countries.

[4] Calculations for Tables 1 and 2 use Hall's procedures.

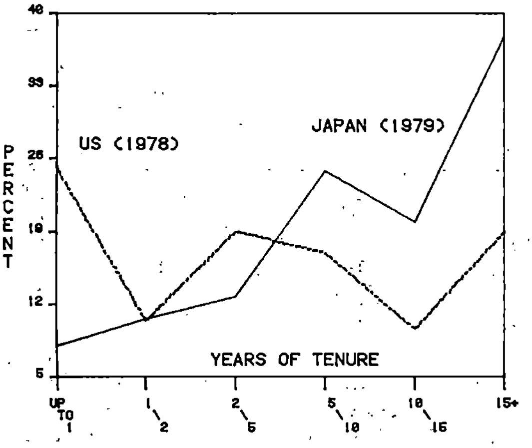


FIGURE 2

quently contain information on employment duration, the necessary data for this calculation fortunately are available for a comparable period in the two countries. Following Hall, the number of workers is divided by the civilian noninstitutional population, rather than by the number of employed persons, to take account of all sources of mobility including mobility between labor-force and non-labor-force activities as well as among different employers.

According to Table 1, for Japan in 1962, 33.5 percent of males ages 15–19 were in the 0–5-year tenure category. Fifteen years later in 1977, 12.2 percent of males ages 30–34 were in the 15–20-year tenure category. (Percent distributions for terminal years, 1977 for Japan and 1978 for the United States, are not reported in Table 1 to save space.) Therefore, the fifteen-year job retention rate for these males is 36.4 percent. As this table clearly illustrates, job retention rates are considerably higher in Japan than in the United States. For example, among those who are 20–24-years old with tenure of 5+ years, 65 percent in Japan retain the same job fifteen years later, but only 30 percent do so in the United States. In both countries, job retention rates are lower among younger workers and among those who are in a low-tenure category at the start. This finding is expected, as young workers are often in the process of shopping for their ultimate employment, while those with low tenure at the

TABLE 1—DISTRIBUTION OF MALE POPULATION BY AGE AND TENURE

Age (1)	Tenure (Years) (2)	Percent of Population (3)	15-Year Retention Rates: Percent With Tenure <1 Yr. (4)	Percent With Tenure >10 Yrs. (5)	1962–77 (%) (6)	Age (7)	Percent With Tenure <10 Yrs. (8)	Percent With Tenure >20 Yrs. (9)
		1962			**Japan**	**1977**		
15–19	0–5	33.5	13.1	0	36.4	30–34	40.0	0.3
20–24	0–5	51.7	8.9	1.0	45.1	35–39	25.2	9.4
	5+	14.4			65.3			
25–34	0–5	27.4	5.0	19.8	42.7	40–49	18.6	30.9
	5+	42.3			73.0			
35–39	0–5	15.7	3.5	37.6	37.7	50–54	16.4	37.5
	5+	49.4			75.9			
		1963			**United States**	**1978**		
14–19	0–5	25.1	16.8	0	5.6	30–34	76.4	0.1
20–24	0–5	64.7	35.9	0.5	13.0	35–39	62.9	1.5
	5+	5.1			30.0			
25–34	0–5	54.3	22.3	9.5	22.2	40–49	48.3	15.6
	5+	32.9			47.3			
35–44	0–5	34.9	14.5	36.5	24.4	50–59	34.1	29.6
	5+	54.3			54.5			
45–49	0–5	28.8	11.7	45.0	17.2	60–64	22.0	23.8
	5+	59.3			40.1			

Sources: Calculated from the *Basic Survey of Employment*, 1962 and 1977 for Japan and the *Special Labor Force Report*, No. 36 (1973) and No. 235 (1978) for the United States.

TABLE 2—THE NUMBER OF DIFFERENT JOBS PER MALE WORKER: JAPAN (1977) AND UNITED STATES (1978)

	New Jobs Started per Year[a]				Cumulative Number of New Jobs to the Age	
	Per Person		Over The Age Interval			
Age	Japan[b] (1)	U.S. (2)	Japan (3)	U.S. (4)	Japan (5)	U.S. (6)
16–19	0.19	0.50	0.76	2.00	0.72	2.00
20–24	0.26	0.48	1.30	2.40	2.06	4.40
25–29	0.13	0.35	0.65	1.75	2.71	6.15
30–34	0.08	0.25	0.40	1.25	3.11	7.40
35–39	0.07	0.18	0.35	0.90	3.46	8.30
40–54	0.05	0.13	0.75	1.95	4.21	10.25
55–64	0.07	0.07	0.70	0.70	4.91	10.95
65–69	—	0.04	—	0.20	—	11.15
70+	—	0.01	—	0.01	—	11.16

Sources: Calculated from the *Basic Survey of Employment* (Japan), 1977, and the *Special Labor Force Reports* (United States), No. 238, 1978.

[a] Cols. 1 and 2 are twice the ratio of persons with new jobs to population, and cols. 3 and 4 are, respectively, cols. 1 and 2 times the age span.

[b] Persons with new jobs in Japan refer to those who changed jobs within a year plus those who were not working a year ago but who are working now. For the U.S., persons with new jobs are those with tenure of 6 months or less.

start are characterized by relatively high turnover.

Columns 4, 5, 8, and 9 in Table 1 provide related information on tenure distributions. Both the proportions with tenure of less than one year in the early 1960's (col. 4) and less than ten years in the late 1970's (col. 8) are smaller in Japan than in the United States. Note also that the proportion of older employees who have worked for the same company for twenty or more years is higher in Japan than in the United States (col. 9). This finding is contrary to the recent claim that such a proportion is greater in Western countries than in Japan.[5] It also reinforces the conclusion drawn from the comparison of Figures 1 and 2 that employment tenures are longer in Japan, and that this pattern is not an artifact of the age-demographic compositions of the respective populations.

Table 2 summarizes our attempt to answer the question: how many jobs will a typical male worker hold during his working life? The appropriate information is contained in columns 5 and 6 and depicted in Figure 3. As described in the notes to the table, this information is determined incrementally using information contained in columns 1–4. For the United States, the number of new jobs per year in column 2 is calculated as twice the reported ratio of population with tenure of six months or less. For Japan, the relevant information is available in terms of the number of persons who changed jobs within the past year plus those who were not working a year ago but are working now. The ratio of this number to population would seem to be a natural counterpart to the U.S. figures. However, there is no way of knowing how often these individuals changed jobs during the year. We simply assume that all these workers have had tenure of six months or less, and double the ratios, as in the U.S. calculations. By doubling the ratio for Japan, we may be overstating the number of new jobs started in that country. Yet, if we did not augment the Japanese magnitudes, the statistics would not be comparable to the U.S. magnitudes, and furthermore would understate the rate of turnover in Japan.[6] Even

[5] See Tajima. Keep in mind that mandatory retirement occurs mostly at age 55 in Japan; thus, the comparison of older workers must be with those under age 55 for Japan.

[6] The U.S. magnitudes are understated to the extent that turnover occurs for individuals who do not hold a job even for six months. This understatement is exacerbated for Japan since the interval length for initial

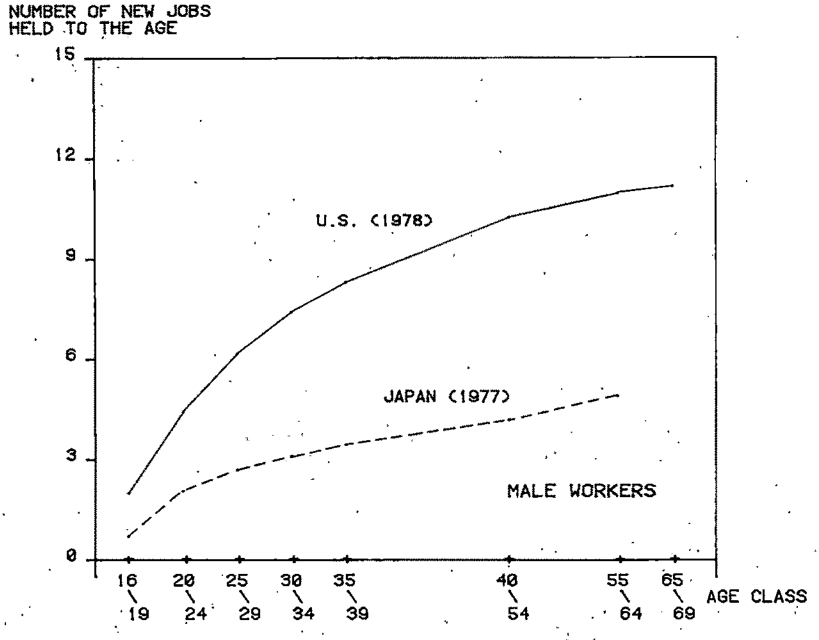


FIGURE 3

with the likelihood of overstatement, we find the cumulative number of jobs held to be much smaller in Japan than in the United States (compare cols. 5, 6, and Figure 3). By the time a typical worker reaches age 65, he would have had approximately five jobs in Japan and eleven jobs in the United States.

According to Table 2, job shopping is more intense for both countries in the early working years. By age 24, the average Japanese male has held 2.06 jobs out of 4.91 jobs he will hold in his working life. An average American male has held 4.4 jobs out of 11 or so jobs he will hold. Interestingly, a typical worker has held a similar percent (40 percent) of the eventual number of jobs by age 24 in both countries.[7] During the next ten years, by age 34, the Japanese male will have held an additional 1.1 jobs, but the American male will have held 3.0 jobs. Close to retirement (age 40–54 in Japan and 55–64 in the United States), 0.05 jobs per year are added in Japan and 0.07 jobs per year in the United States. It is apparent from this table that job turnover during the life cycle is much greater in the United States than in Japan.[8]

With regard to firm-size differences, Japanese scholars have often emphasized what

tenure is one year. We feel it is important to make every attempt to have the statistics between the two countries be comparable, so we doubled the Japanese magnitudes. In our judgement, this procedure overstates the true comparable Japanese magnitudes, but it has the effect of making the contrasts between the two countries to be lower-bound estimates of the true contrasts, and serves our purposes well.

[7]Note that we are talking about the life cycle pattern of job changes in this table. American youths generally change jobs more often than youths in other industrialized countries. In the early 1970's, the percent of youths between the ages of 18 and 24 with work experience who had changed employers more than 4 times was a striking 27.6 percent for the United States, a figure which was much higher than those for France (17.4), England (14.3), Sweden (10.8), Switzerland (10.5), West Germany (5.7) and Japan (2.7). (See Robert Cole, 1979, p. 95.)

[8]Our calculations for females also indicate that U.S. women change jobs more than Japanese women (results available on request). It should be noted that labor turnover in Japan was not always low. Indeed, turnover in the early 1900's was quite high (Koji Taira, 1970, ch. 6), and the industrial relations were far from being harmonious during early post-World War II years (Haruo Shimada, 1983).

TABLE 3—PERCENT OF EMPLOYED MALES BY TENURE AND FIRM SIZE: 1979

	Japan					United States				
Years	All (1)	Tiny (2)	Small (3)	Medium (4)	Large (5)	All (1)	Tiny (2)	Small (3)	Medium (4)	Large (5)
<1	7.9	10.4	9.7	7.7	3.8	19.4	29.7	22.3	16.0	11.2
1–4	21.5	24.8	22.2	21.6	13.5	30.6	37.7	37.1	32.2	24.4
5–9	22.2	20.4	22.4	23.9	21.8	19.4	16.0	18.5	21.3	21.7
10–14	17.7	16.4	16.6	18.9	19.1	11.7	7.6	8.9	12.6	15.4
15–19	11.9	9.1	10.0	12.3	16.2	6.5	3.9	5.2	7.3	8.4
20+	18.8	18.7	16.0	15.7	25.6	12.4	5.0	7.9	10.5	18.8
Median	8.2	8.0	8.0	8.1	12.0	4	2	3	5	7
Eventual Tenure	25.0	23.6	22.4	23.6	30.8	15.6	9.6	12.2	15.4	20.6

Sources: Calculated from the *Basic Survey of Employment* (Japan), 1979, and the *Current Population Survey* (United States), 1979.
Note: Employed males in private industries. Tiny = 1–9 for Japan and 1–25 for United States, Small = 10–99 for Japan and 26–99 for United States, Medium = 100–999, and Large means 1000+ for both countries. Eventual tenure is calculated as twice the mean tenure.

they consider to be the unique role of firm size in the Japanese economy, as evidenced in the literature on the "dualistic" labor market (see, for example, Konosuke Odaka, 1967; Yasukichi Yasuba, 1976; Carl Mosk and Yoshi Nakata, 1983). In the United States, however, firm size has not been central to the study of wages and employment (for exceptions, see Richard Lester, 1967; Stanley Masters, 1969; Albert Rees and George Schultz, 1970). Recently, however, perhaps stimulated by the availability of firm-size information in the recent supplement to the *Current Population Survey*, there has been renewed interest in firm-size differences in wage and employment characteristics (see Wesley Mellow, 1982; Walter Oi, 1983a,b). We focus at this point on how job tenure differs by firm size, as reported in Table 3, and offer additional information on firm-size differences in other economic magnitudes in Table 4. The latter table is noteworthy because firm-size differences in other selected variables are quite similar in the two countries. Thus, rather than being unique to Japan, the so-called dualistic market phenomenon may reflect little more than some underlying causal factors which operate in the two countries, or any country for that matter.

In Table 3, we present tenure distributions by four firm-size categories for each of the countries. Also, we present the median tenure and a crude estimate of the eventual tenure of jobs that are currently held. The latter statistic is computed by doubling the mean tenure value, a procedure that is common in the unemployment duration literature (see Stephen Salant, 1977; George Akerloff and Brian Main, 1981). All of these statistics indicate that the employer-employee attachment is stronger in Japan than in the United States. There is also a definite tendency for longer tenures to occur in larger firms. One reason for firm-size differences in job tenure is the well-known higher failure rate for small firms than for large firms. If the differential failure rate were the sole factor, however, the compensating principle would lead one to predict higher wages in smaller firms. Available evidence suggests that the opposite is true. For example, Haruo Shimada has shown that, in both countries, wages are higher in larger firms throughout workers' life cycles (1981, pp. 71–74). Also, according to Robert Cole's study of workers in Yokohama (1979, pp. 87–90), the proportion of job leavers due to involuntary discharge, which presumably includes the case of business failures, is, if anything, highest in the largest firms and lowest in the smallest firms. Thus, there must be more to firm-size differences in job tenure than just the differential failure rates.[9] Failure

[9]Another reason for lower turnover in larger firms may be that workers in these firms have more opportunities to change jobs without changing employers. Indeed, according to Cole's evidence (pp. 80–81), intrafirm mobility increases with firm size in both Yokohama and Detroit.

TABLE 4—FIRM-SIZE DIFFERENCES[a]

	All Industries				Manufacturing				Nonmanufacturing			
	Small (1)	Med. (2)	Large (3)	(3/1)	Small (1)	Med. (2)	Large (3)	(3/1)	Small (1)	Med. (2)	Large (3)	(3/1)
Earnings[c]						Japan 1980						
(a) Actual	196.7	214.9	257.3	1.3	195.6	213.9	251.2	1.3	197.3	215.6	263.1	1.3
(b) Contract	181.6	193.0	223.7	1.2	176.8	187.1	212.1	1.2	184.1	197.0	234.7	1.3
(c) Bonus	482.3	731.2	1069.7	2.2	480.6	739.3	980.4	2.0	483.2	725.7	1154.3	2.4
Ratio: c/a	2.5	3.4	4.2		2.5	3.5	3.9		2.4	3.4	4.4	
c/b	2.7	3.8	4.8		2.7	4.0	4.6		2.6	3.7	4.9	
Hours Worked												
(a) Contract	194	182	167	0.9	195	180	165	0.8	193	183	169	0.9
(b) Overtime	15	18	21	1.4	19	21	23	1.2	13	16	19	1.5
Ratio: b/a	0.08	0.10	0.13		0.10	0.12	0.14		0.07	0.09	0.11	
Education[b] (Percent)												
Low	41.2	29.7	26.1	0.6	52.3	37.3	33.5	0.6	35.4	24.5	19.1	0.5
College	11.9	22.2	25.7	2.2	8.4	17.4	18.0	2.1	13.8	25.5	33.0	2.4
Tenure (Years)	8.5	10.2	13.9	1.6	9.5	11.6	14.7	1.5	8.0	9.4	13.1	1.6
Age	39.0	37.0	37.1	1.0	40.4	37.2	36.8	0.9	38.3	36.9	37.4	1.0
Earnings ($)						United States 1979						
Annual	13445	16715	18500	1.4	13883	15882	18712	1.3	13357	17103	18284	1.4
Weekly	267.5	329.5	359.6	1.3	270.2	306.3	361.6	1.3	267.0	340.3	357.6	1.3
Hours Worked												
Weeks/Yr.	49.8	50.7	51.3	1.0	51.0	51.7	51.7	1.0	49.5	50.2	51.0	1.0
Hours/Wk.	42.0	42.8	42.4	1.0	41.7	42.5	42.6	1.0	42.0	42.9	42.2	1.0
Education[b] (percent)												
Low	27.9	19.0	15.7	0.6	34.2	27.9	19.5	0.6	26.7	14.8	11.9	0.4
College	16.3	29.4	27.0	1.7	9.2	15.4	18.9	2.1	17.7	35.9	35.1	2.0
Tenure (Years)	5.3	7.7	10.3	1.9	6.0	8.8	11.6	1.9	5.1	7.2	9.0	1.8
Age	36.2	37.6	38.4	1.1	36.8	38.3	39.0	1.1	36.1	37.3	37.8	1.0
Union (Percent)	16.3	33.2	44.7	2.7	19.1	34.4	51.8	2.7	15.7	32.6	37.6	2.4

[a]Employed males in nonagricultural private industries.
[b]Low means less than 12 years of schooling, college means 16 or more years of schooling.
[c]1000 Yen: Actual and Contract Earnings are Monthly, Bonus is Annual.

rates are relevant, however, insofar as they affect the incentive to invest in firm-specific capital.

Although job tenure is longer in large Japanese firms, it is quite long even in the tiny and small firms. This finding raises doubts about the popular belief that life-time employment is only a large-firm phenomenon. Indeed, some recent studies indicate that there exists a significant degree of "paternalism" among small-scale Japanese employers.[10] Moreover, Cole's evidence noted above casts doubt on the common observation that firing in Japan is done mostly by small firms. Our evidence together with Cole's suggests that small Japanese firms managing to survive do exhibit a tendency for long-term employment. Thus, the common assertion that Japanese small firms operate much like casual labor markets is not entirely accurate.

To summarize, long-term employment relationships are more prevalent in Japan than in the United States, though both our evidence and that of Hall's indicate that employer-employee attachment is by no means

[10]In an illuminating ethnographic study of a small pencil maker in a low-income section of Tokyo, George De Vos and Hiroshi Wagatsuma (1973), describe how small firms try to emulate the traditional "paternalism" employment relationship. Employees are let go only in the event of business failure or antisocial behavior. Yasuba (pp. 285–86) argues that the traditional paternalistic and personalistic human relations continued into employer paternalism in small firms during the early years of Japan's modernization. Cole distinguishes between the "personalized paternalism" of small firms and the "administrative paternalism of the large firm" (pp. 21–22). Admittedly, the concept of paternalism is not economic, but we find these references significant insofar as they underscore the existence of long-term employment relationships even among small firms.

weak in the United States. In both countries, one observes a tendency for larger firms to have a stronger employment relationship. Contrary to popular belief, however, employment tenure in small Japanese firms is not universally short.

II. Earnings Profiles in Japan and the United States

We now investigate the difference in earnings profiles between the two countries. Although some evidence is available on this issue (for example, Masatoshi Kuratani, 1973; Shimada, 1981), it is useful to reexamine the evidence for two reasons. First, not only are our data more recent than the data used in the previous studies, but they enable us to control for firm-size effects directly, a feature lacking in the previous studies. Second, previous studies relied solely on the regression coefficient of tenure on earnings, or the slope of the earnings profile, to draw conclusions. For reasons discussed below, a proper and fuller comparison is based instead on the number of years it takes to reach the peak earnings and the value of the peak earnings as well as the slope.

The recent literature offers a variety of reasons for expecting the earnings-tenure profile to be related to employment duration. Notable explanations include the firm-specific human capital, the screening, and the agency hypotheses.[11] The literature is unsettled as to the relative merits of these hypotheses, however. To assess which hypothesis dominates as an explanation of the differences between the Japanese and the U.S. labor markets is a formidable task, and is beyond the scope of this paper. Rather, the purpose of this section is to determine if Japan-U.S. differences in employment tenure and the shape of earnings profiles are consistent with the prediction, common to these hypotheses, that jobs with longer employment duration will have more steeply sloped tenure profiles.

In specifying an earnings function, two key variables are experience acquired in the current firm (i.e., tenure) and previous experience. The sum of the two experience components is the total, or potential work experience. We estimate Mincer-type earnings regressions, which hold constant total experience (see Jacob Mincer, 1974, and Mincer and Boyan Jovanovic, 1981). In particular, we examine coefficient estimates of the following earnings function for each of the two countries:

$$(1) \qquad \ln Y = a_0 + a_1 j + a_2 j^2 + b_1 n + b_2 n^2 + cjn + zd + e,$$

where $\ln Y$ is the logarithm of earnings, j is the total years of work experience (i.e., age minus years of schooling minus 6), n is the number of years of tenure in the current firm, z is a vector of other explanatory variables, e is the disturbance term, the a's, b's, and c are constants, and d is a vector of constants.

The human capital literature distinguishes general and firm-specific capital components. The empirical specification in equation (1) accommodates this distinction, but interpretation of the coefficients on tenure variables as solely reflecting specific human capital effects does not necessarily follow. These coefficients could reflect the combined impacts of a number of competing hypotheses, notably screening and agency. Inclusion of the total experience magnitudes in the specification in our view does not misrepresent other hypotheses. Both screening and agency hypotheses are silent regarding the role of total experience when analyzing the earnings-tenure relationship. However, neither hypothesis rules out the existence of positive effects of total work experience on earnings, that is, general human capital effects.

We are concerned primarily with the pure effects of tenure on earnings net of total experience effects. The functional form of

[11] The specific-capital hypothesis is by now familiar. See Gary Becker (1962), Hashimoto (1979), and Lorne Carmichael (1983). According to Joanne Salop and Steven Salop (1976), employers may use upward-sloping earnings profiles to discourage those with high propensity to quit from seeking employment (the screening hypothesis). Edward Lazear (1979) offers the agency hypothesis, which views steeply sloped earnings profiles as a device to discourage employee shirking.

equation (1) allows for interaction between the two experience profiles: slopes of either profile obviously depend on the amounts of both experience magnitudes, since $\partial \ln Y/\partial j = a_1 + 2a_2 j + cn$ and $\partial \ln Y/\partial n = b_1 + 2b_2 n + cj$. We investigated more elaborate interaction schemes, but found little difference in the qualitative findings. Therefore, this paper reports results using the simplified interaction scheme.

The dependent variables are the logarithms of monthly earnings including bonuses for Japan, and of usual weekly earnings for the United States. As for the explanatory variables, the total years of experience is estimated conventionally as age minus years of schooling minus 6. For Japan, we use dummy variables for schooling; the data are reported by category rather than years. High school means 12–13 years, junior college means 14–15 years, and university means 16+ years of schooling. The excluded category is less than 12 years of education. For the United States, educational attainment is measured in years, but we also include two dummy variables, high school for those who have at least 12 years of education, and university for those who have 16 or more years. These dummy variables are meant to capture the effects of completing either high school or college. Finally, union member is a dummy variable to distinguish union from nonunion members. This information is available for the United States sample only.

The data used to estimate equation (1) are from the *Basic Survey of Wage Structure* (Shugyo Kozo Kihon Chosa, 1980) for Japan and the public use tapes of the *Current Population Survey* (*CPS*, March and May, 1979) for the United States, and are for male workers in nonagricultural private industries. The Japanese data are cell means of earnings, bonus payments, years of tenure, and age of worker, cross classified by firm size, education class, age class, and industry. The *CPS* data contain individual observations, and the May 1979 *CPS* includes information on firm size, a feature which is particularly useful for this research.

Table 5 reports the *OLS* regression estimates of equation (1) for Japan and the United States, separately for three firm-size groups. Small firms are those employing fewer than 100 workers, medium firms, between 100 and 999, and large firms, more than 1000 workers. Both total experience and tenure variables are statistically significant at better than the 1 percent level in all cases.[12] These coefficients generate the usual concave earnings profiles for both total experience and tenure profiles.[13] The interaction of total experience and tenure is negative for Japan, though significant only for small firms, but is positive for the United States with high statistical significance for medium and large firms. These interaction coefficients suggest that, in Japan, previous years of experience, a component of total experience, tend to penalize the earnings growth due to tenure. In contrast, in the United States, previous years of experience boost the tenure effects on earnings. This difference in the interaction effects in the two countries may be related to our earlier findings of job changes being more frequent in the United States than in Japan.[14]

[12]The *F*-values for the significance tests of total experience are 280.4, 143.3, and 8.1 for Japanese small, medium, and large firms, respectively, and 180.3, 82.0, and 94.5 for the U.S. small, medium and large firms, respectively. The *F*-values for the tenure variables are 53.7, 29.4, and 53.7 for Japanese small, medium, and large firms, respectively, and 27.6, 9.6, and 38.2 for the U.S. small, medium and large firms, respectively.

[13]Because of the interactive earnings specifications, to generate a *ceteris paribus* profile for tenure requires that total experience be fixed at some level. The same is true for the total experience profile. We checked the profile properties at respective mean values of total experience and tenure. The means for Japanese small, medium, and large firms, respectively are: for total experience, 22.3, 19.4, and 19.4 years; for tenure, 8.5, 10.2, and 14.6 years. For the U.S. small, medium and large firms, respectively, the means are: for total experience, 18.1, 18.5, and 19.2 years; for tenure, 5.3, 7.7, and 10.3 years.

[14]In the U.S. regressions, the union coefficient is positive and significant only for small firms while it is insignificant for medium and large firms. Since no interactions are specified between the union and the occupation and/or education variables in these regressions, one must use caution in interpreting the union coefficient. It certainly would be premature to infer that unions have little effect on wages in medium and large firms. Workers in larger firms have higher levels of education, and highly educated union members are less likely than others to have been promoted in the past, a likelihood which lowers their wages. Support for this conjecture is

TABLE 5—REGRESSIONS OF MALE EARNINGS FOR JAPAN AND THE UNITED STATES[a]

	Japan Firm Size (1980)			U.S. Firm Size (1979)		
	Small (1)	Medium (2)	Large (3)	Small (4)	Medium (5)	Large (6)
Constant	4.6163	4.6519	4.7213	4.3879	4.6152	4.791
	(284.5)	(265.6)	(180.8)	(68.3)	(53.1)	(78.1)
Total Experience:						
j	0.0377	0.0523	0.0210	0.0502	0.0477	0.0372
	(15.4)	(18.8)	(4.9)	(22.2)	(15.6)	(16.4)
j^2	−0.0005	−0.0008	−0.0003	−0.0010	−0.0010	−0.0007
	(−12.4)	(−18.9)	(−4.1)	(−21.5)	(−14.4)	(−13.0)
Tenure:						
n	0.0614	0.0278	0.0692	0.0196	−0.0019	0.0121
	(9.1)	(5.2)	(11.7)	(4.8)	(−0.4)	(4.3)
n^2	−0.0003	−0.0001	−0.0013	−0.0006	−0.0003	−0.0003
	(−0.8)	(−0.5)	(−5.6)	(−4.0)	(−2.1)	(−3.6)
Interaction: $j \times n$	−0.0012	−0.0002	−0.00001	0.0003	0.0006	0.0003
	(−6.1)	(−1.9)	(−0.1)	(1.8)	(3.3)	(2.3)
Years of Schooling	–	–	–	0.0367	0.0362	0.0296
				(5.6)	(4.2)	(5.0)
Schooling Dummies[b]						
High School	0.1278	0.1368	0.1048	0.1630	0.1735	0.1286
	(15.2)	(12.7)	(5.8)	(4.9)	(3.9)	(4.5)
Junior College	0.2552	0.2635	0.2822	–	–	–
	(12.9)	(12.1)	(7.2)			
University	0.3700	0.3923	0.4579	0.0630	0.0808	0.1590
	(28.9)	(29.3)	(20.8)	(1.7)	(1.8)	(5.6)
Union Member	–	–	–	0.2533	0.0099	−0.0020
				(10.0)	(0.4)	(−0.1)
MSE	c	c	c	0.2626	0.1804	0.1590
$\bar{R}^2$	c	c	c	0.316	0.307	0.279
N	384	378	366	3139	1590	3750

Sources: Basic Survey of Wage Structure (Shugyo Kozo Kihon Chosa), 1980, and the *Current Population Survey*, March and May 1979.

[a]The t-values are shown in parentheses; the dependent variable = logarithm of earnings in private nonagricultural industries.

[b]Schooling dummies indicate: For Japan, High School = 12–13 years, Junior College = 14–15 years, and University = 16+ years of schooling; and for the United States, High School is for those who have at least 12 years of education and University for those who have 16+ years.

[c]Since Japanese estimates are derived from weighted regressions, these statistics are not readily available as summary statistics in the computer output. Furthermore, when grouped data are used, these statistics are not comparable to those obtained by using data on individual observations.

What do these regressions indicate about the effects of firm-specific tenure on earnings? The effects of total experience and tenure are captured by the slopes of the two profiles. The slopes change, however, as one moves along each of the profiles, and their comparisons across firm-size groups or between the two countries will be affected by the choice of the point on the profiles where such comparisons are made.[15] A proper

provided in the samples broken down by education. Union coefficients are significantly positive regardless of firm size in regressions in the sample of those with less than high school education. In contrast, union coefficients are negative, and significant for medium and large firms, in regressions using those with 16+ years of education. These results are available upon request. Finally, the coefficients for the schooling variables contain few surprises.

[15]As a result, care must be taken when interpreting the slope as a meaningful indicator. For example, slope is commonly used as an indicator of investments in human capital. In particular, a greater slope would not necessarily indicate greater investment opportunities. Imagine two earnings profiles, A and B, where A has a steeper slope than B, but A reaches its peak sooner and

TABLE 6—PERCENT GROWTH IN EARNINGS ATTRIBUTABLE TO EXPERIENCE AND TENURE SINCE ENTERING THE CURRENT FIRM, BY FIRM SIZE

Tenure (Years)	Small Firms		Medium Firms		Large Firms	
	Total	Tenure	Total	Tenure	Total	Tenure
Japan						
5	56.0	36.8	44.7	17.5	50.7	40.5
15	180.2	123.0	151.9	70.3	166.9	140.0
20	222.6	148.6	203.1	99.2	213.9	181.0
25	235.5	147.8	242.9	123.4	239.8	203.2
30	215.3	117.9	264.7	138.1	238.8	200.5
35	167.7	65.2	264.7	140.3	210.9	173.5
Peak[a]	235.6	150.4	267.4	141.0	242.8	205.2
	(24)		(33)		(27)	
Starting Pay[b] (Monthly Yen)	115,935		121,185		129,663	
United States						
5	36.9	11.7	23.3	< 0	25.3	7.1
15	108.4	40.7	67.2	4.1	74.4	26.1
20	131.5	52.2	84.0	10.2	93.0	36.4
25	140.0	57.9	94.8	18.7	105.4	45.6
30	131.9	56.7	98.6	28.9	109.7	52.6
35	109.1	49.0	94.8	39.3	105.6	56.7
Peak[a]	140.0	57.9	98.6	28.9	109.7	52.6
	(25)		(30)		(30)	
Starting Pay[b] (Weekly $)	160.83		187.90		201.61	

Notes: Calculated from Table 5. These figures represent the percentage differences between the current earnings and the starting pay due to total experience and tenure. Total experience effects are derived from $d \ln Y/dj$ as indicated in fn. 16.

[a]The figures indicate the peak growth rates due to total experience (peak years are in parentheses) and the associated growth rates due to tenure.

[b]The starting pay is estimated at zero experience and tenure and the mean values of explanatory variables.

comparison requires an examination of the number of years it takes to reach the peak earnings and the value of the peak earnings itself as well as the slope of the earnings profile.

In view of these considerations, we report in Table 6 what the estimated regression coefficients imply for the percentage growth rates in earnings between the year when a worker joins the current firm and various years of tenure. In constructing the table, we asked the following question: if a worker enters the current firm right after completing schooling and stays with the firm until he retires, by how much would his earnings grow as he accumulated years of tenure? We calculate separately the growth rates attributable to total experience and to firm-specific tenure. The table reports both total and firm-specific earnings growth, although our primary interest is in the latter.[16] In addition, Table 6 reports the peak values of the total growth percentages, the years when they occur, the associated growth due to firm-specific tenure, and the starting pay. We think that the most meaningful country and firm-size comparisons of growth are made between the initial and the peak years. The starting pay is estimated by evaluating equa-

at a lower peak value than B. In this case, B may easily reflect greater ultimate investments than A, though its slope is smaller than that for A at a particular tenure value.

[16]The total earnings growth due to the accumulation of total experience can be decomposed as follows: $d \ln Y/dj = \partial \ln Y/\partial j + \partial \ln Y/\partial n$, since we are assuming $dn/dj = 1$, or that the worker stays with the firm throughout his working life. The two partials can be evaluated by using the estimated regression coefficients.

tion (1) at zero values of j and n, and the means of the explanatory variables.

A worker entering a small Japanese firm enjoys a 56.0 percent growth in his earnings by the fifth year, at which time a 36.8 percentage point growth is associated with firm-specific tenure. He reaches his peak growth of 235.6 percent in the twenty-fourth year, at which time a 150.4 percentage point growth is generated by firm-specific tenure.[17] An eye-opening pattern in this table concerns the Japan-U.S. differences in the growth rates due to firm-specific tenure as well as total experience. For all firm-size groups, growth rates between the peak earnings year and the initial year are greater in Japan than in the United States. Note also that within firm-size groups, peak years are similar in the two countries.[18] Thus, it is slopes of earnings profiles that are driving our comparisons between Japan and the United States. More importantly, growth rates attributable to tenure are far greater in Japan than in the United States.[19]

Do the observed Japan-U.S. differences in earnings profiles reflect the heterogeneity of the sample in the two countries? Worker quality and the industrial composition, for example, are different between the two countries.[20] To what extent are our findings upheld once these differences are netted out? To shed light on this question, we performed the analysis by disaggregating the data by industry and education. As it turns out, Japan-U.S. differences in the earnings-tenure profile persist in both manufacturing and nonmanufacturing industries and within education class.[21]

According to the findings in this section, both the earnings-tenure and earnings-total experience profiles are more steeply sloped in Japan than in the United States, holding constant firm size. These findings together with the finding in Section I of a longer employment tenure in Japan than in the United States are consistent with the existing theories of employment contracts.

III. Summary and Conclusion

Contrary to the impression created by the recent literature, Japanese male workers do have longer employment tenure than American workers. Also, Japanese workers have steeper earnings profiles that peak in about the same year after entering the current firm as American workers. In particular, growth rates in earnings attributable to tenure are far greater in Japan than in the United States. These findings are robust to disaggregation by industrial composition and educa-

[17] In the human capital perspective, these figures imply somewhat larger ratios of specific capital to total growth rates than those found by Mincer and Jovanovic (pp. 38–42). The main reason for the difference between their estimates and ours is that we are performing a simulation for a hypothetical individual who stays with the firm throughout his working life (see fn. 13), whereas they evaluate these ratios at a point in time implicitly taking into account the probabilities of separation.

[18] Firm-size differences in the percentage growth due to firm-specific tenure exhibit a mixed pattern, however. In Japan, workers in large firms experience the greatest growth, but the difference between workers in small and medium firms is negligible. The differences between small and large firms, as well as between medium and large firms, are statistically significant at the 1 percent level (t-values are 2.75 and 3.99, respectively). Thus, large firms appear to be distinct from either medium or small firms. In the United States, the relationship between the percentage growth and firm size appears U shaped, with the growth rate somewhat higher in small firms than large firms. However, the only statistically significant difference is between medium and large firms with a t-value of 2.16.

[19] Table 6 is based on workers who begin working in the current firm immediately after completing schooling. A more realistic case may be where workers join the current firm after working in other firms for some years. We have examined cases with various years of previous experience (results available upon request) and conclude that the findings in Table 6 are robust.

[20] In our samples, workers in service and trade industries predominate in U.S. small firms to a greater extent than in Japan. The percent of small-firm workers in these industries is 51 percent for the United States and 37 percent for Japan. Comparable figures for large firms are 27 percent for the United States and 16 percent for Japan. For the difference in worker quality, as measured by education, see Table 4.

[21] These results are not reported to save space but are available upon request. Also, our general findings resemble those reported in an interesting book by Shimada (1981), which uses the data for earlier years. However, his regression specifications differ from ours, and we are able to control for firm size directly in the U.S. data whereas Shimada was unable to do so with his data.

tional attainment. The pattern of Japan-U.S. differences in the employment tenure and the earnings profiles are consistent with three prominent hypotheses in the literature—the specific human capital, the screening, and the agency hypotheses.

It is beyond the scope of this paper to investigate the relative contribution of each of the hypotheses to the differences we uncovered between the two countries. However, a few remarks may be in order on the screening and the agency hypotheses, which are the recent challenge to the traditional specific-capital hypothesis. It is worth noting that the screening hypothesis does not necessarily deny the existence of specific capital. Indeed, according to a behavioral version of this hypothesis, employers' desire to minimize the loss of specific capital caused by turnover is the principal motive for screening potential employees (see Salop and Salop). The agency hypothesis, however, predicts upward-sloping earnings profiles even if specific capital is nonexistent. Among the two behavioral hypotheses, therefore, we consider the agency hypothesis to be the more challenging alternative to the specific-capital hypothesis.

In view of the above consideration, we conclude this paper by considering briefly how our evidence bears on the comparison of the agency with the specific-capital hypothesis. According to the agency hypothesis, workers in effect post bonds in early years for their performance, and the bonds are greater the higher the monitoring costs. The slope of the earnings-tenure profile may be viewed as indicating the extent of bonding. Our findings indicate that Japanese workers post larger bonds than American workers. Is this finding consistent with the agency hypothesis? The answer depends in part on the relative employer-employee transaction costs in the two countries. Hashimoto (1979) has analyzed the Japanese practice of bonus payments by arguing that such transaction costs are lower in Japan than in the United States.[22] If this argument is valid, bonds should be smaller in Japan than in the United States, which is contrary to what we find. Our evidence then would seem to contradict the agency hypothesis, if the transaction-cost assumption is valid. Such a conclusion is premature, however, and a definitive verdict must await further evidence on the relative transaction costs in the two countries.[23]

The pattern of Japan-U.S. differences in employment tenure and earnings-tenure profiles is consistent with there being more specific human capital in Japan than in the United States. The specific-capital hypothesis, however, predicts mandatory retirement to occur at a later age in Japan than in the United States. In fact, mandatory retirement typically occurs earlier in Japan (between ages 55 and 60) than the United States, an uncomfortable fact to those who subscribe to the specific-capital hypothesis. Since the setting of a mandatory retirement age reflects a multitude of factors, including life expectancy and worker productivity, what is relevant perhaps is the change in, rather than the level of, the retirement age. It is interesting to note that, during the past ten years or so, a number of employers have raised their mandatory retirement age above the traditional age of 55, though most retirement still occurs below 60. In view of this, it would seem premature to understate the relative importance of the specific capital hypothesis in explaining Japan-U.S. labor market differences. To develop a strong case for this hypothesis, however, one needs to study the factors responsible for the difference and change in Japanese and U.S. retirement practices. Such a study would further our

[22] In the bonus system, workers put enormous trust in management, which announces the profit conditions of the firm. Why would workers accept such a payment scheme in which a large portion of their annual earnings depends on management's word unless the costs of transaction were low? The bonus system increases wage flexibility. Wage flexibility is by no means unique to Japan, however. Although bonus-type arrangements tend to be confined in the United States to management level workers, recent evidence suggests that nonmanagement workers also experience wage flexibility (see Raisian, 1983).

[23] See our 1984 working paper for some evidence supporting this assumption.

understanding of the determinants of Japan-U.S. differences in employment tenure and earnings profiles.

REFERENCES

Akerloff, George A. and Main, Brian G. M., "An Experience-Weighted Measure of Employment and Unemployment Duration," *American Economic Review*, December 1981, *71*, 1003–11.

Becker, Gary S., "Investment in Human Capital: A Theoretical Analysis," *Journal of Political Economy*, October 1962, Suppl., *70*, 9–49.

Bruce-Briggs, B., "Lifetime Employment: A Non-Lesson from Japan," *Wall Street Journal*, August 1982.

Carmichael, Lorne, "Firm-Specific Human Capital and Promotion Ladders," *Bell Journal of Economics*, Spring 1983, *14*, 251–58.

Cole, Robert E., *Work, Mobility, and Participation*, Berkley: University of California Press, 1979.

De Vos, George and Wagatsuma, Hiroshi, "The Entrepreneurial Mentality of Low-Class Urban Japanese in Manufacturing Industries," in George De Vos, ed., *Socialization for Achievement: Essays in the Cultural Psychology of the Japanese*, Berkley: University of California Press, 1973, 201–19.

Hall, Robert E., "The Importance of Lifetime Jobs in the U.S. Economy," *American Economic Review*, September 1982, *72*, 716–24.

Hashimoto, Masanori, "Bonus Payments, On-the-Job Training and Lifetime Employment in Japan," *Journal of Political Economy*, October 1979, *87*, 1086–104.

______ **and Raisian, John,** "Employment Tenure and On-The-Job Training in Japan and the United States," Working Paper, No. E-84-4, Hoover Institution, 1984.

Koike, Kazuo, *Shokuba no Rodo Kumiai to Sanka* (Trade Union and Participation on the Shop Floor), Tokyo: Tokyo Keizai Shimpo Sha, 1977.

______, "Japan's Industrial Relations: Characteristics and Problems," *Japanese Economic Studies*, Fall 1978, *7*, 42–90.

Kuratani, Masatoshi, "The Theory of Training and Employment: An Application to Japan," unpublished doctoral dissertation, Columbia University, 1973.

Lazear, Edward P. "Why is There Mandatory Retirement?," *Journal of Political Economy*, December 1979, *87*, 1261–84.

Lester, Richard, "Pay Differentials by Size of Establishment," *Industrial Relations*, October 1967, *7*, 57–66.

Masters, Stanley H., "An Interindustry Analysis of Wages and Plant Size," *Review of Economics and Statistics*, August 1969, *51*, 341–45.

Mellow, Wesley, "Employer Size and Wages," *Review of Economics and Statistics*, August 1982, *64*, 495–501.

Mincer, Jacob, *Schooling, Experience, and Earnings*, New York: National Bureau of Economic Research, 1974, 21–63.

______ **and Jovanovic, Boyan,** "Labor Mobility and Wages," in Sherwin Rosen, ed., *Studies in Labor Markets*, Chicago: University of Chicago Press, 1981.

Mosk, Carl and Nakata, Yoshi, "Dualism and *Nenko*: Age Structure and Wages in Japan, 1961–81," Working Paper No. 171, University of California-Berkeley, 1983.

Odaka, Konosuke, "The Structure of Japanese Labor Markets," *Kikan Riron Keizaigaku* (*Economic Studies Quarterly*), June 1967, *18*, 25–42.

Oi, Walter Y., (1983a) "Heterogeneous Firms and the Organization of Production," *Economic Inquiry*, April 1983, *21*, 147–71.

______, (1983b) "The Fixed Employment Costs of Specialized Labor," in Jack E. Triplett, Ed., *The Measurement of Labor Cost*, Chicago: University of Chicago Press, 1983.

Raisian, John, "Contracts, Job Experience and Cyclical Labor Market Adjustments," *Journal of Labor Economics*, April 1983, *2*, 152–70.

Rees, Albert and Schultz, George, *Workers and Wages in Urban Labor Market*, Chicago: University of Chicago Press, 1970.

Salant, Stephen W., "Search Theory and Duration Data: A Theory of Sorts," *Quarterly Journal of Economics*, February 1977, *91*, 39–57.

Salop, Joanne and Salop, Steven, "Self-Selection and Turnover in the Labor Market,"

Quarterly Journal of Economics, November 1976, *90*, 619–27.

Shimada, Haruo, *Earnings Structure and Human Investment: A Comparison Between the United States and Japan*, Tokyo: Keio Economic Observatory, 1981.

______, "New Challenges for Japanese Labor Management Relations in the Era of Global Structural Change," *Japan Labor Bulletin*, July 1983, *22*, 5–8.

Taira, Koji, *Economic Development and the Labor Market in Japan*, New York: Columbia University Press, 1970.

Tajima, Mitsuo, "Japan's Cold Indifference Towards Old People," *Wall Street Journal*, November 1982.

Umemura, Mataji, "The Seniority-Merit Wage System in Japan," in S. Nishikawa, ed., *The Labor Market in Japan: Selected Readings*, Tokyo: University of Tokyo Press, 1980, 177–87.

Yasuba, Yasukichi, "The Evolution of Dualistic Wage Structure," in Hugh Patrick, ed., *Japanese Industrialization and Its Social Consequences*, Berkeley: University of California Press, 1976, 249–98.

Japan, Labor Minister's Statistical Information Office, *Basic Survey of Wage Structure* (Shugyo Kozo Kihon Chosa), 1980.

______, **Prime Minister's Statistical Office,** *Basic Survey of Employment*, 1962; 1977.

U.S. Bureau of the Census, *Current Population Survey*, March; May, 1979.

U.S. Bureau of Labor Statistics, Special Labor Force Report No. 36, Washington: USGPO, 1973.

______, Special Labor Force Report Nos. 235 and 238, Washington: USGPO, 1978.

Notches

By ALAN S. BLINDER AND HARVEY S. ROSEN*

Tax and transfer systems provide numerous incentives that influence people's behavior. Sometimes, these incentives are inadvertent by-products of the need to raise revenue. Other times, public policy deliberately aims to change behavior. In either case, the approach typically favored by economists is to change the slope of some relevant budget constraint by introducing a tax or a subsidy. A Pigouvian emissions fee is a clear example.

Economists have an instinctively negative reaction to any program that creates a "notch"; that is, a discontinuity in a budget constraint. Perhaps the best known example of a notch comes in the welfare system, where such programs as Medicaid and public housing are structured so that a finite lump of benefits is lost all at once when a household's income crosses a certain threshold. The reason for economists' negative attitudes toward this notch is clear: for people with low earnings potential, the notch effectively imposes a very high marginal tax rate over a small income range, which no doubt discourages work and promotes welfare dependency.

Such notches deserve their bad reputations. However, this paper argues that, in other contexts, tax and subsidy plans with notches may have been dismissed too cavalierly, and should at least be considered as serious contenders when public policy seeks to encourage or discourage some activity. Since this idea is so foreign to our normal way of thinking, perhaps we should develop the intuition behind it at the outset.[1]

A standard tax or subsidy alters the relative price that everyone faces, and hence distorts everyone's behavior. A notch, on the other hand, leaves the effective price unchanged—except, of course, at the notch, where the price is undefined. Consequently, a standard tax or subsidy imposes small excess burdens on everyone, while a notch imposes large excess burdens on a small number of people. Stated this way, it is not immediately obvious that the notch approach is always inferior. Indeed, this paper produces several examples in which notches are clearly superior.

Although notches have a bad name among economists, they are not uncommon in the private sector. Some airlines stimulate the demand for air travel not by lowering the price per ticket, but by offering a free ticket for passengers who have flown more than a certain number of miles. Similarly, banks and savings and loan associations occasionally attempt to increase deposits not by offering a higher rate of interest, but by awarding a "gift" to customers who deposit an amount exceeding some specified level. This paper argues that notches may sometimes be appropriate in the public sector as well.

The paper is organized as follows. Section I uses a simple example—the tax deductibility of charitable contributions—to explain the basic ideas, develop a methodology for addressing the issue, and then compare the relative efficacy of notches vs. traditional lin-

*Department of Economics, Princeton University, Princeton, NJ 08544. We are grateful to Avner Bar-Ilan for assistance with the computations and for helpful suggestions. We also thank Rebecca Blank, David Bradford, Angus Deaton, Roger Gordon, Peter Hartley, Jerry Hausman, the referees, and participants in seminar presentations at Princeton, Yale, and MIT for useful comments. This research was supported by the National Science Foundation and the Taxation Program of the National Bureau of Economic Research.

[1] The optimal tax literature has discussed general conditions under which nonlinear taxation is more efficient than linear taxation. See Jesus Seade (1977). A notch incentive can be viewed as a special case of a nonlinear tax which is sufficiently simple that it is a viable policy option. The two-part tariff is another well-known example of how departures from linear pricing can enhance welfare.

ear subsidies in stimulating charitable giving. We conclude that a notch incentive may dominate a traditional linear incentive. Some practical problems involved in implementing notch incentives are discussed in Section II, and Section III contains concluding remarks.

I. Stimulating Demand for a Commodity

Suppose the government wants to stimulate one person's charitable giving, which we denote in Figure 1 by F (for "favored" commodity). If the person's marginal tax rate is t, deductibility of charitable contributions lowers the effective price of each dollar of charity from \$1 to \$$(1-t)$, thereby pivoting the budget constraint between F and all other goods from MN out to MO. As a consequence, the individual's optimum moves from E_1 to E_2; charitable giving winds up at F_2.

The government could induce an identical increase in charity by the following notch system: if the individual donates F_2 or more, he receives a lump sum subsidy equal to DE_2; otherwise, he receives nothing. The budget constraint associated with this notch scheme is MDE_2H, and the optimal choice remains E_2. Thus the notch and linear schemes have exactly the same revenue cost, DE_2, and induce the same behavior.

This example illustrates an obvious point. As long as only one individual is being considered—or equivalently, if it is possible to design a separate notch scheme for each individual—then there is nothing to choose between a linear incentive and a notch incentive. However, in the realistic case in which individuals have heterogeneous tastes, and all taxpayers (at least within the same income class) face the same budget constraint, then the notch and linear subsidy schemes can have quite different implications.

To see why, consider Figure 2. Under the status quo, individuals A, B, and C all face budget constraint MN. Individual A's highest attainable indifference curve is labeled U_A, and similarly for B and C. Now suppose that a notch subsidy of G, equal to distance KD in the figure, is granted to anyone who donates at least F^*. The budget line confronting all three individuals becomes $MDKI$. As seen in the diagram, individual A's behavior is unchanged. Individual C moves up to indifference curve U_C'. Note that the movement from U_C to U_C' induced by the subsidy creates no excess burden—it is equivalent to a lump sum grant. Now consider individual B. His best choice under the notch subsidy is right at the notch (indifference curve U_B'). In contrast to C, the subsidy to B does create an excess burden. Individual B would be better off with a lump sum subsidy of KD, which would produce budget constraint HI, allowing B a utility level U_B'' which exceeds U_B'.

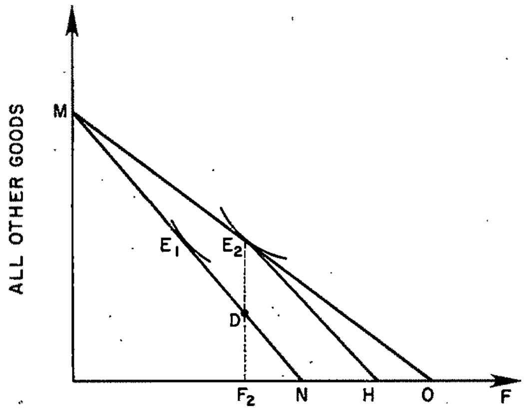


FIGURE 1

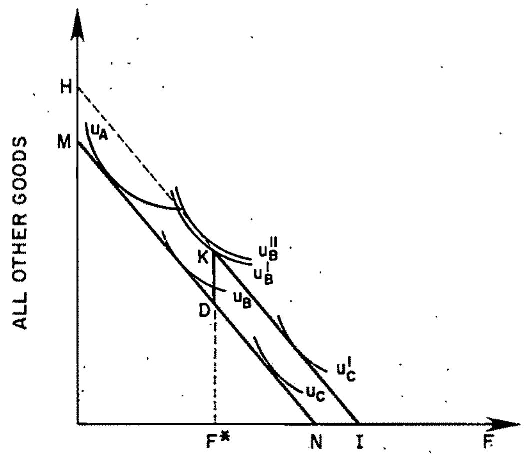


FIGURE 2

Thus, as we stated in the introduction, under a notch scheme, some individuals face no excess burdens, and some face large ones. The aggregate excess burden depends on the distribution of individual tastes. Generally, the more people that behave like B, the higher will be the total excess burden under a notch. The revenue costs of notch subsidies similarly depend on the distribution of tastes.

Can it be that the notch scheme is preferable to a conventional linear subsidy? Because heterogeneity of the population is the essence of the matter, it is difficult to obtain an analytical answer even for very simple assumptions about utility functions and the joint distribution of their parameters. For this reason, we rely on simulation techniques. Our strategy is to posit specific utility functions and numerical distributions of their parameters, and investigate the relative desirability of notch and linear incentives under alternative assumptions.

A. *The Basic Setup*

Let an individual's utility level U depend on his consumption of a composite commodity Y, whose price is equal to unity, and F, the favored commodity. Suppose the utility function is of the constant elasticity of substitution form:

$$(1)\qquad U=\left[aF^{-b}+(1-a)Y^{-b}\right]^{-1/b},\qquad b>-1,$$

where a and b are taste parameters, and the elasticity of substitution is given by $\sigma=1/(1+b)$. With a linear subsidy at rate s, the individual's budget constraint is

$$(2)\qquad (1-s)P_F F+Y=M+Z,$$

where P_F is the price of the favored good, M is income, and Z is any lump sum transfers (positive or negative) from the government.

Maximizing utility subject to constraint (2) leads to the following demand functions:

$$(3)\qquad F=\frac{M+Z}{(1-s)P_F+K\left((1-s)P_F\right)^{\sigma}},$$

$$(4)\qquad Y=\frac{K(M+Z)}{K+\left((1-s)P_F\right)^{1-\sigma}},$$

where $K=[(1-a)/a]^{\sigma}$. Note that since the utility function is homothetic, the Engels' curves are rays through the origin as long as Z is proportional to M. Thus, any results that apply to one income level can be scaled up or down to apply to other income levels.

It will be useful to have a formula for the indirect utility function, which turns out to be:

$$(5)\qquad V=\left[(M+Z)/Q(s)\right]\left[A(s)\right]^{-1/b},$$

$$\text{where}\quad A(s)=a+(1-a)\left[K(1-s)^{\sigma}P_F^{\sigma}\right]^{-b}$$

$$\text{and}\quad Q(s)=(1-s)P_F+K\left[(1-s)P_F\right]^{\sigma}.$$

B. *Analysis of a Linear Subsidy*

We turn first to the efficiency and revenue consequences of a linear subsidy. To compute the subsidy's excess burden, we must find the monetary value of the difference between the utility level achieved with the subsidy and the level that would have been achieved had the subsidy been given as a lump sum. Thus, note that the cost to the treasury for a given individual is sFP_F. Suppose that sFP_F had been given to the individual as a lump sum, without distorting relative prices. Then, using equation (5), utility would be

$$(6)\qquad V_0=\left(\frac{M+sFP_F}{Q(0)}\right)\left[A(0)\right]^{-1/b}.$$

If V_L is utility with the linear subsidy, then we define the excess burden of the subsidy as the amount of money we would have to take away from the individual at V_0 to lower his utility to V_L.[2] Algebraically, the excess

[2] There are a number of other ways to define excess burden, but for our purposes the differences are inconsequential. See Alan Auerbach and Rosen (1980).

burden, B_L, is implicitly defined by

$$[M/Q(s)]A(s)^{-1/b} = [(M+sFP_F-B_L)/Q(0)]A(0)^{-1/b},$$

which has the closed-form solution

$$B_L = M\left[1+\frac{sP_F}{Q(s)}-\frac{Q(0)}{Q(s)}\left(\frac{A(s)}{A(0)}\right)^{-1/b}\right].$$

The aggregate excess burden of the linear subsidy is simply the sum of B_L across individuals, and the total cost to the treasury is the sum of sP_FF.

C. *Analysis of a Notch Subsidy*

Now consider a scheme which awards a lump sum subsidy of G to individuals who donate at least some critical amount F^*, and zero otherwise. In calculating the excess burden of the notch subsidy, three possibilities must be considered.

(a) The individual's optimal decision is unchanged. In this case, there is no excess burden and no budgetary cost. This case corresponds to individual A in Figure 2.

(b) The individual is induced to consume right at the notch. Here the cost of the subsidy is clearly G per individual. There is also an excess burden, which corresponds to the difference between utility levels U_B'' and U_B' in Figure 2. To find the individual's utility level at the notch, we must evaluate the direct utility function[3] at $F=F^*$ and $Y=M+G-P_FF^*$, yielding

$$V_N = \left[a(F^*)^{-b} + (1-a)(M+G-P_FF^*)^{-b}\right]^{-1/b}.$$

Following the reasoning behind equation (6), the utility if the same subsidy had been granted as a lump sum is

$$[(M+G)/Q(0)]A(0)^{-1/b}.$$

Hence, the excess burden for such an individual, B_N, is implicitly defined by

$$a(F^*)^{-b}+(1-a)(M+G-P_FF^*)^{-b} = [(M+G-B_N)/Q(0)]^{-b}A(0)$$

or explicitly by

$$(7)\quad B_N = M+G-Q(0)\times \left\{\frac{a(F^*)^{-b}+(1-a)(M+G-P_FF^*)^{-b}}{A(0)}\right\}^{-\frac{1}{b}}$$

(c) The individual consumes more than the critical quantity F^* after the notch scheme is imposed. In this case, the notch subsidy is equivalent to a lump sum transfer, so again there is no excess burden. But there is a cost to the treasury, namely G. This corresponds to individual C in Figure 2.

To summarize: if there are n_a individuals in category (a), n_b in category (b), and n_c in category (c), then the total cost of the notch subsidy is $(n_b+n_c)G$, and the total excess burden is the sum of B_N defined by equation (7) across individuals in category (b).

D. *Simulation Strategy*

The simulations assume a population of 499 people. Individuals are indexed by their value of a, the share parameter in the utility function. The values of a are distributed uniformly between 0.0002 and 0.0998.[4] Within each simulation, the elasticity of substitution is the same for all individuals. But its value is varied across simulations. The units of F are chosen to make P_F, its before-tax price, equal to one. Each person's income M is also normalized to 1. Thus the simulations should be thought of as applying to a given income class. Since our utility func-

[3]The indirect utility function (5) cannot be used because of the discontinuity in the slope of the budget constraint at F^*.

[4]That there are 499 individuals rather than 500 is a quirk of the simulation. Originally, we had let the parameter a run between .002 and .998, and were forced to omit both endpoints in order to get an interior solution. When we moved the decimal point on a, we neglected to include $a=.10$.

TABLE 1—SUBSIDY TO CONSUMPTION OF A COMMODITY WITH COBB-DOUGLAS UTILITY FUNCTIONS UNIFORM DISTRIBUTION OF a

			Notch Subsidy		
	Status Quo (1)	Linear Subsidy $s = 0.2$ (2)	(a) $F^* = .03$ $G = .164$ (3)	(b) $F^* = 0.183$ $G = 0.035$ (4)	(c) $F^* = 0.116$ $G = 0.015$ (5)
Total Demand for F	24.3	30.5	30.5	30.5	30.5
Total Revenue Cost	0	6.1	81.9	2.4	2.6
Total Excess Burden	0	0.62	1.04	2.0	1.19
TEC					
$(m = 0.2)$	0	1.8	17.4	2.5	1.7
$(m = 0.4)$	0	3.1	33.8	3.0	2.2

tions are homothetic, this is not a substantive restriction.

We first compute aggregate demand for F in the absence of any subsidies. Then we impose a linear subsidy of 20 percent ($s = 0.2$), and compute the amount by which consumption of F is stimulated, the revenue cost to the government, and the total excess burden. Next, we turn to the notch subsidy, and search over various combinations of G and F^* to find those that yield the same total demand for F. In general, an infinite number of notch schemes are consistent with any fixed value of aggregate F. For example, if F^* is set very high, but at the same time G is large, it might be possible to achieve the same aggregate value of F as when both parameters are low. Criteria for choosing among the various notch schemes are discussed below.

E. *A Cobb-Douglas Result*

We begin by discussing the Cobb-Douglas case.[5] Column 1 of Table 1 shows that the aggregate consumption of the favored good in the absence of any subsidy is 24.3 units. Column 2 indicates that imposition of a 20 percent subsidy increases the quantity demanded to 30.5.[6] The cost to the treasury of the subsidy is 6.1, and the associated excess burden is 0.62. Thus the subsidy is relatively efficient in this case: the deadweight loss per dollar of subsidy is only about 10 cents.[7] This is worth pointing out, because it shows that the simulation is set up to make the linear subsidy scheme hard to beat.

Our next goal is to devise a notch scheme that induces the same change in behavior, and compute its revenue cost and excess burden. As just noted, an infinite number of notch schemes can do the trick. Columns 3, 4, and 5 compare three possibilities.

In column 3 (notch subsidy (a)), the notch is set at a relatively low level of consumption, $F^* = .03$ units, which is about one-half the mean consumption level under the linear subsidy. With F^* set this low, we must induce essentially everyone in the population to donate at least F^* in order to reach the target for aggregate donations. And, to accomplish this, a large grant of $G = .164$ (16.4 percent of total income) is required. The revenue costs of setting F^* at such a low level are revealed in the second row: the cost to the treasury is 81.9, about thirteen times greater than the cost of the linear subsidy. The excess burden of 1.04 also far exceeds that of the linear subsidy. It is clear that notch subsidy (a) is a perfectly dreadful idea.

Under notch subsidy (b) in column 4, the notch is set at a very high value, 0.183, or 18.3 percent of income. This is about three times the average consumption level under

[5]Actually, for computational reasons, our "Cobb-Douglas" case is $\sigma = 1.01$.

[6]Note that the product of price times quantity is approximately unchanged (i.e., $30.5 \times 0.8 \doteq 24.3$) because of the Cobb-Douglas utility function.

[7]To cite just one comparison, John Weicher (1979) presents estimates that each $1 spent on public housing yields about 85 cents worth of benefits to recipients.

the linear subsidy. With required consumption so high, most people do not take advantage of the subsidy; only 15 percent receive the grant of 0.035. As a consequence, the revenue cost is far lower than under the linear subsidy—only 2.4. However, those people who *do* accept the subsidy have their behavior distorted considerably; the excess burden is 2.0—more than triple the excess burden under the linear subsidy.

A final possibility, notch subsidy (c), is exhibited in column 5 of Table 1. The notch is placed at a lower level than in column 4 but a higher level than in column 3: $F^* = .116$. The revenue cost is higher than in column 4, but the excess burden is lower. This is because the people whose behavior changes are nearer the notch, and hence their decisions are less distorted.

The results in Table 1 taken together suggest that while some notch incentives (such as notch subsidy (a)) will do quite horribly compared to linear schemes, others will do quite well. Both schemes (b) and (c), although they have higher excess burdens than the linear subsidy, have much lower revenue costs. On balance, therefore, they might be preferable.

This observation leads to an important question. Given that we are judging subsidy systems on the basis of *two* criteria, revenue cost and excess burden, how are we to compare them when one is better on one criterion and the second on the other? There are two possibilities:

1) The most natural approach is to compare the schemes on the basis of what we call *total efficiency cost* (*TEC*), defined as the sum of the excess burden arising from distorting the demand for the favored good plus the efficiency cost of raising the revenue needed to finance the subsidy. If taxes were lump sum, the efficiency cost of replacing the revenue lost by the subsidy would be zero, and the *TEC* would just be the standard excess burden of the subsidy. However, real world tax finance creates its own efficiency costs. If m is the marginal excess burden created by a dollar of taxes raised in the private sector, then the *TEC* is just m times the revenue loss plus the excess burden from distorting the consumption of F.

What is the value of m? The answer depends on what tax instrument the government uses, and the supply and demand elasticities of the item(s) being taxed. Traditionally, the marginal excess burden of taxation has been supposed to be quite low. Recent estimates are higher. Charles Ballard, John Shoven, and John Whalley (1982) estimate that the marginal excess burden of \$1 raised via the corporate tax is about \$0.50. This is the same as the estimate obtained by Charles Stuart (1984) for the whole tax system. Jerry Hausman's (1981) econometric study of labor supply suggests that if \$1 is raised by a tax on labor income, the marginal excess burden is about \$0.42. In the absence of agreement on what the marginal excess burden of taxation is, it makes no sense to restrict ourselves to one figure. We therefore do calculations assuming values of m of both 0.2 and 0.4 which, if anything, seem on the low side.[8]

The *TEC* figures are recorded in the last two rows of Table 1. Assuming $m = 0.2$, notch subsidy (b) with a *TEC* of 2.5, is inferior to the linear subsidy, whose score is 1.8. However, notch subsidy (c), with a *TEC* of 1.7, is better than the linear subsidy. In fact, when $m = 0.2$, notch subsidy (c) is the most efficient of all possible notch subsidies in the sense of having the minimum *TEC*. When $m = 0.4$, both notch schemes (b) and (c) are better than the linear subsidy.

These results suggest that the outcome depends critically on the assumed value of m. The reason is clear. Reasonable notch schemes make smaller demands on the treasury than linear schemes, but often have larger excess burdens. If m is high, a great deal of weight is given to the smaller revenue cost, thereby enhancing the attractiveness of notch schemes. If m is low, the excess burden is relatively more important—which enhances the attractiveness of linear schemes.

2) The second way to compare various subsidy programs questions the relevance of excess burden in this context. After all, the whole exercise of measuring excess burden assumes that the subsidy "distorts" behavior

[8]Since our tax changes are not infinitesimal, we should really use the average marginal rate over the relevant range, which would presumably exceed m.

TABLE 2—SUBSIDY TO CONSUMPTION OF A COMMODITY WITH ALTERNATIVE ELASTICITIES OF SUBSTITUTION UNIFORM DISTRIBUTION OF a

	$\sigma=0.5$			$\sigma=1.5$		
	Status Quo (1)	Linear Subsidy $s=0.2$ (2)	Notch Subsidy $F=0.28$ $G=0.03$ (3)	Status Quo (4)	Linear Subsidy $s=0.2$ (5)	Notch Subsidy $F=.04$ $G=.007$ (6)
Total Demand for F	87.4	99.7	99.7	6.90	9.62	9.62
Total Revenue Cost	0	19.9	6.4	0	1.92	0.95
Total Excess Burden	0	.92	2.80	0	0.29	0.51
TEC						
($m=0.2$)	0	4.90	4.09	0	0.68	0.70
($m=0.4$)	0	8.89	5.37	0	1.06	0.89

away from the optimum. This might, of course, be the case if the subsidy was instituted solely in response to political pressures. On the other hand, it is possible that the subsidy is deliberately put in place to correct an externality.[9] As is well known, a subsidy levied on a good that generates positive externalities might actually enhance efficiency; that is, have a negative excess burden. In this case, the standard excess burden calculation does not really make sense.

This discussion leads to the following conclusion. In some instances, it may make more sense simply to look at the budgetary cost of achieving the required increase in consumption of the favored good, making no allowance for excess burden. And a glance at Table 1 indicates that if minimizing revenue cost is the sole criterion, a well-chosen notch scheme may well be superior to a linear subsidy.[10] By targetting the subsidy to those whose tastes for the favored commodity are relatively intense, the notch subsidy does not "waste" money on those whose consumption is not stimulated very much.

On the basis of Table 1, then, we conclude that notch subsidies may be better than linear subsidies. This is more likely to be true when the distortion that arises is not considered to generate a deadweight burden, or when a relatively high weight is put on the efficiency costs of financing revenue losses. But it can also be true in other cases.

F. *Changing the Elasticity of Substitution*

How does the relative attractiveness of linear and notch subsidies depend upon the elasticity of substitution between the favored commodity and all other goods? To answer this question, we repeat the simulation assuming $\sigma=0.5$ and $\sigma=1.5$. The results are reported in Table 2. In each case, we

(i) compute consumption of the favored good in the absence of any subsidy (cols. 1 and 4);

(ii) compute consumption, revenue cost, and excess burden associated with a 20 percent linear subsidy (cols. 2 and 5);

(iii) use numerical methods to find the best notch subsidy that induces the same amount of consumption as the linear subsidy, where "best" means that the *TEC* is at a minimum for $m=0.2$.[11]

When $\sigma=0.5$, the notch subsidy has an excess burden about three times that of the

[9]It is far from obvious that this is the rationale for government subsidization of charitable giving.

[10]This discussion assumes that only the aggregate quantity of the externality-producing commodity matters. In the case of classical externalities, efficiency requires that *each person's* consumption of F be such that the social marginal benefit equals the marginal cost. Just like standard Pigouvian subsidies, our scheme ignores the fact that the optimum corrective subsidy or tax generally differs across individuals.

[11]We also report results for $m=0.4$. The characteristics of the notch schemes that minimize *TEC* for $m=0.4$ are not very different from those that minimize it for $m=0.2$. The minimum was found by using a simple grid search.

linear subsidy, but a revenue cost less than one-third as large. This pattern is already familiar from the Cobb-Douglas case examined in Table 1. If we consider revenue costs alone, the notch scheme is certainly superior by a wide margin. If we consider both excess burden and efficiency costs, and assume $m = 0.2$, we find that the *TEC* for the notch scheme (4.09) is about 17 percent smaller than that of the linear scheme (4.90). In the Cobb-Douglas case of Table 1, the comparable improvement was only about 8 percent. This suggests that the attractiveness of the notch subsidy is enhanced when the elasticity of substitution decreases.

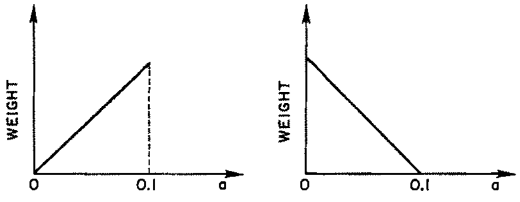
This impression is confirmed when we examine the results for $\sigma = 1.5$ reported in the right side of Table 2. When $\sigma = 1.5$, the total efficiency cost of the notch subsidy when $m = 0.2$ slightly *exceeds* that for the linear subsidy. To be sure, the notch scheme is still much cheaper, but this is not enough to counter its deficiency on the excess burden criterion when $m = 0.2$. However, when $m = 0.4$, the notch scheme is preferred to the linear scheme even with $\sigma = 1.5$.

The explanation for these results lies in the fact that low values of σ translate into low price elasticities of demand for F. Thus, relatively large values of the linear subsidy rate are required in order to achieve any given change in the demand for F. For example, Table 2 shows that with $\sigma = 0.5$, a 20 percent subsidy raises consumption only 14 percent. Because the linear subsidy works entirely via price effects, low elasticities of substitution make the linear subsidy more expensive.

On the other hand, the notch subsidy achieves much of its stimulus through income effects and, with a *CES* utility function, the income elasticity of demand is unity regardless of the value of the elasticity of substitution. (See equation (3).) Thus, notch subsidies are more attractive in the presence of low values of the elasticity of substitution, *ceteris paribus*.

G. *Changing the Distribution of a*

So far we have been assuming that the share parameter, a, is distributed uniformly over the interval $(0, .1)$. We now consider



(a) Population of F-Lovers (b) Population of F-Haters

FIGURE 3

some simple nonuniform linear distributions. Specifically, we continue to assume that there are a total of 499 people and that a runs between .0002 and .0998, but we make the density of a skewed. Figure 3a shows the assumed distribution of a for a population with more F-lovers than F-haters; Figure 3b shows the opposite case. Our analyses of nonuniform distributions assume Cobb-Douglas utility functions and these linear densities.

Results are reported in Table 3. As before, for each configuration of utility function parameters, we present the equilibrium under the assumption of no subsidy, a linear subsidy of 20 percent, and the most efficient notch subsidy that induces the same behavioral change as the linear subsidy.

Regardless of whether the population is comprised mostly of F-lovers or F-haters, the notch subsidy has much lower revenue costs, but much higher excess burdens than the linear subsidy. However, in all cases the notch subsidy is preferred on the basis of *TEC*—albeit by a trivial margin in one case.

Note that on the basis of *relative TEC*, the notch subsidy is more preferred when the population is heavily weighted toward F-lovers. When $m = 0.2$ and the population has more F-haters, the *TEC* of the notch scheme is 98 percent of the linear subsidy's. But when the population has more F-lovers, the comparable figure is 82 percent.[12] Why? The fundamental advantage of a notch scheme is that it targets the subsidy to the "right" people. This advantage is worth most when

[12] Observe from Table 1 that the comparable figure for a uniform distribution lies between these two.

TABLE 3—SUBSIDY TO CONSUMPTION OF A COMMODITY WITH COBB-DOUGLAS UTILITY FUNCTIONS
LINEAR DISTRIBUTION OF a

	More F-Haters			More F-Lovers		
	Status Quo (1)	Linear Subsidy $s=0.2$ (2)	Notch Subsidy $F=0.110$ $G=0.018$ (3)	Status Quo (4)	Linear Subsidy $s=0.2$ (5)	Notch Subsidy $F=0.114$ $G=0.012$ (6)
Total Demand for F	16.2	20.3	20.3	32.5	40.7	40.7
Total Revenue Loss	0	4.05	1.78	0	8.13	3.20
Total Excess Burden	0	0.42	0.97	0	0.82	1.38
TEC						
$(m=0.2)$	0	1.23	1.22	0	2.45	2.02
$(m=0.4)$	0	2.04	1.69	0	4.07	2.66

many people have intense preferences for the commodity. With a lot of the population near the notch, few people have to be "dragged" very far in order to reach the notch, and therefore the excess burden per unit of induced consumption is relatively small.

H. *Summary*

We have compared notch and linear subsidies for stimulating consumption of a favored commodity, such as charitable giving. In every case we have examined, a notch subsidy can stimulate the same increase in consumption at a lower cost to the treasury. Hence, if revenue loss is the only criterion, the notch subsidy is clearly superior. However, if the increase in demand for the favored commodity induced by the subsidy is viewed as a distortion, then the associated excess burden must also be taken into account. Here, notch schemes seem to do systematically worse than linear subsidies. However, when the revenue costs and excess burdens are suitably aggregated to find a measure of total efficiency cost, the notch scheme often does better.[13]

In this context, it is important to remember that we did not set out to establish that notch schemes are generally, or even typically, superior to linear schemes. Our purpose was only to show that notch schemes might be superior under circumstances that are in no sense pathological. This seems to have been shown.

[13] Presumably, a two-parameter subsidy scheme that combines a notch with a conventional linear subsidy would be superior to either pure system. Indeed, this is the case. But when we experimented with such a mixed system, we found that the optimal two-parameter scheme combined a very large notch with a linear *tax*.

II. Some Practical Problems

Three major practical problems would arise in attempts to implement notch incentives:

1. *Differences across income classes.* Our simulations examine populations with identical endowments. With a homothetic utility function, once a notch incentive that "works" for one income class is found, the same scheme can be scaled upward or downward and applied to every class. If people's utility functions are not homothetic, then the analysis must be done separately for each income group. Moreover, if policymakers wish to achieve certain distributional goals at the same time that behavior is being modified, then, as usual, efficiency criteria alone cannot be used to compare various plans. An explicit social welfare function with distributional weights must be introduced. Neither of these problems is "new" to the notch approach; they both arise in the linear case as well.

2. *Bunching (intertemporal substitution).* Like all nonlinear tax and transfer plans, notch subsidies provide an incentive to bunch the subsidized activity into particular time periods. For example, if a $100 grant is given in

any year in which charitable giving reaches $1000, then individuals who would otherwise give $500 per year might instead give zero and $1000 in alternate years.

The seriousness of this problem depends on the context. In the case of charitable giving, it might be quite serious; bunching is therefore a formidable obstacle to the implementation of a notch subsidy. In other contexts, it might be less serious (see Section III).

Where intertemporal substitution is a severe problem, the remedy is obvious: lifetime averaging, which would make the lifetime, rather than the year, the relevant unit of time. In the case of charitable giving, averaging seems quite feasible. After all, the gift tax in the United States is now handled precisely in this way. In other applications, averaging may be more difficult administratively.

In any case, the principle is clear: notches will not look very attractive in applications where the elasticity of intertemporal substitution is high and/or where lifetime averaging is difficult.

3. *Cheating* (*interpersonal substitution*). Taxpayers might collude in order to obtain notch subsidies or avoid notch taxes. If a grant of $100 is awarded to anyone who gives $1000 or more in a year, then two individuals who donate $500 each are ineligible. However, if one turns over his receipts to the other (or lets the other do his donating), the latter can claim the $100 subsidy, and then split it with his collaborator.

Is this likely to be a serious problem? It depends upon the transactions costs of such collaboration, and these will vary from case to case. In the case of transferring charitable contributions, these might be fairly low. (However, the answer depends in part on the importance that people attach to having *their* name associated with *their* contribution.) In contrast, cheating in other contexts seems less likely (see again Section III).

III. Conclusions

We have analyzed the consequences of notch incentives—taxes and subsidies that create jumps in budget constraints—as opposed to linear incentives, which simply change the slopes of budget constraints. Unlike linear incentives, notch schemes do not distort the behavior of every person. Rather, if properly designed, they induce individuals to self-select so that those who are most willing to change their behavior are the ones who receive the subsidy (or avoid the tax). In the cases we have examined, notch schemes do not uniformly dominate linear schemes. However, they come out on top often enough that they deserve serious consideration as policy options.

For purposes of exposition we have concentrated on the issue of charitable giving. However, a number of other possible applications exist:

a) *Housing*. Under current law, homeowners receive substantial tax deductions through the personal income tax. Ostensibly, the purpose of these provisions is to stimulate homeownership, and they seem to have been effective in increasing the number of homeowners. But, at the same time, the subsidy increases the amount of housing purchased by homeowners (see Rosen, 1979). A notch scheme targetted just at homeownership—a lump sum reduction of tax liability for those who own homes—could be designed to have the same effect on the homeownership rate, with less severe revenue and efficiency consequences. Since the subsidy for homeownership would presumably apply only to one house per family, neither intertemporal nor interpersonal substitution should create problems.[14,15]

b) *Education*. Currently, tuition tax credits are being considered as a way to promote higher education. A notch version of this scheme could give a family a lump sum payment only after (say) a certain number of years of college education had been

[14] Australia currently awards a lump sum subsidy to first-time purchasers of houses. We thank Brian Wright for informing us of this.

[15] Interestingly, a federally sponsored social experiment conducted several years ago investigated a notch subsidy to stimulate the consumption of *rental* housing by the poor. Each family in the experiment received a lump sum grant if its monthly housing expenditure exceeded some critical amount. See Steven Venti and David Wise (1984) for details.

bought. Since each child in the household could receive the subsidy only once, the bunching issue would not arise. And cheating would require falsification of college records. Hence neither problem seems important in this context.

c) *Saving*. There are several provisions in the tax code designed to encourage saving, and others have been proposed. Generally, these are linear subsidies that raise the return to saving, perhaps up to some maximum amount.[16] Consider as an alternative a notch subsidy that offers a lump sum grant to those who save more than a certain amount. (The threshold amount would obviously be keyed to income.) We have done some extensive simulations with such plans, and found that they often (but not always) are a better way to stimulate saving than replacing the income tax by a consumption tax.

The drawback is a practical one: intertemporal substitution is a potentially devastating problem in this case. Individuals can easily bunch their saving into particular years so as to avail themselves of the notch subsidy without really saving more in the long run. Indeed, the problem is worse than that because savings eligible for the subsidy would presumably have to be deposited into particular accounts (analogous to IRAs). Individuals could easily transfer funds into and out of these accounts in order to give the appearance of bunching their saving even though, in fact, they were saving at a smooth rate. In brief, the problem of intertemporal substitution may preclude the use of notch subsidies to encourage saving.

d) *Welfare Reform*. As mentioned in the introduction, our current welfare system is characterized by notches that provide strong disincentives to work. In particular, poor families lose substantial benefits such as Medicaid and public housing when their earnings cross certain thresholds. Economists have strongly condemned the current system for its adverse incentives. Many have advocated replacing it by a negative income tax (*NIT*) which would, in essence, get rid of the notches. Both a priori reasoning and empirical evidence support the idea that an *NIT* is a better way to redistribute income than our current welfare system.

We, naturally, join the condemnation of the present system and agree that an *NIT* would be better. But we believe that the problem with the current system is not so much that it employs a notch as that it employs a *perverse* notch. It is possible that a notch welfare system would provide positive work incentives for the working poor.

Specifically, consider the following sort of welfare plan. Individuals are offered a lump sum grant, G, on the condition that they work at least a certain number of hours, H^*, per year. If they work less than H^* hours, they receive nothing.[17] In a set of simulations not reported here, we compared the current welfare system, an *NIT*, and a notch scheme. The notch scheme not only dominated the *status quo*, but for certain configurations of the parameters, it stimulated more labor supply with less excess burden than the *NIT*.

Interestingly, neither intertemporal nor interpersonal substitution would likely be a problem for a notch welfare scheme of the sort we propose. Low-wage people have limited opportunities to vary their hours of work, and a year is already a long time period as such decisions go. Similarly, the costs of cheating are probably very substantial in the welfare case. Presumably, worker A would have to get his employer to report some of his earnings as if they had been earned by worker B. This requires the complicity of the employer. Then worker B would have to be trusted to share the subsidy check that he receives from the government. The whole thing sounds quite cumbersome. These

[16]David Bradford (1980) provides a survey of these provisions.

[17]The system we outline provides nothing for households unable to work. Thus it must be viewed as one component of a categorical system that also includes an outright dole for those judged unable to work. Of course, the administrative difficulties posed by categorical systems is one reason why many reformers favor the *NIT*. It is far from clear, however, that any real world welfare system can be entirely noncategorical. It is hard to imagine the political process allowing a healthy full-time college student to receive *NIT* payments, for example.

and other considerations lead us to suspect that cheating might not be a major problem for a welfare system that featured a notch incentive to work.

We have emphasized throughout that the results depend on the distribution of individual tastes and endowments. If the ideas advanced here are deemed to be fruitful, the natural next step would be to simulate the effect of notch subsidies for realistic programs using estimated utility function parameters for a sample of actual consumers. Such a study would provide a more definitive basis for evaluating the efficacy of notch incentives. All we have done here is to drop a few, hopefully provocative, hints.

REFERENCES

Auerbach, Alan J. and Rosen, Harvey S., "Will the Real Excess Burden Please Stand Up? (Or, Seven Measures in Search of a Concept)," Working Paper No. 495, National Bureau of Economic Research, June 1980.

Ballard, Charles L., Shoven, John B. and Whalley, John, "The Welfare Costs of Distortions in the United States Tax System: A General Equilibrium Approach," Working Paper No. 1043, National Bureau of Economic Research, December 1982.

Bradford, David, "The Economics of Tax Policy Towards Savings," in George M. von Furstenberg, ed., *The Government and Capital Formation*, Cambridge: Ballinger, 1980.

Hausman, Jerry A., "Income and Payroll Tax Policy and Labor Supply," in Laurence Meyer, ed., *The Supply Side-Effects of Economic Policy*, St. Louis Federal Reserve Bank, 1981.

Rosen, Harvey S., "Housing Decisions and the U.S. Income Tax: An Econometric Analysis," *Journal of Public Economics*, February 1979, *11*, 1–23.

Seade, Jesus, "On the Shape of Optimal Tax Schedules," *Journal of Public Economics*, April 1977, *7*, 203–35.

Stuart, Charles E., "Welfare Costs per Dollar of Additional Tax Revenue in the United States," *American Economic Review*, June 1984, *74*, 352–62.

Venti, Steven F. and Wise, David A., "Moving and Housing Expenditure: Transactions Cost and Disequilibrium," *Journal of Public Economics*, February/March 1984, *23*, 207–43.

Weicher, John C., "Urban Housing Policy," in P. Mieszkowski and M. Straszheim, eds., *Current Issues in Urban Economics*, Baltimore: Johns Hopkins Press, 1979.

Searching for Leviathan: An Empirical Study

By WALLACE E. OATES*

> Total government intrusion into the economy should be smaller, *ceteris paribus*, the greater the extent to which taxes and expenditures are decentralized....
>
> [*Brennan and Buchanan*, 1980, p. 185]

In several papers and a recent book, Geoffrey Brennan and James Buchanan (1977, 1978, 1980) have put forth a striking and controversial view of the public sector. Drawing by analogy on the conventional theory of monopoly in the private sector, they envision a monolithic government that systematically seeks to exploit its citizenry through the maximization of the tax revenues that it extracts from the economy. From this perspective, they develop a fiscal constitution whose central purpose is to constrain "Leviathan" by limiting in various ways its access to tax and other fiscal instruments.

While the Leviathan hypothesis has been the source of lively debate and a wide range of policy proposals, it has not been the subject of much systematic empirical work or testing.[1] This is a matter of some importance since the policy implications of the Leviathan view are disturbing, to put it mildly. In particular, Brennan and Buchanan virtually stand on their heads many of the basic theorems in public finance for an efficient and equitable tax system. If, in fact, the Leviathan view is an inaccurate depiction of the functioning of the public sector, the introduction of their policy proposals is likely to make a sorry mess of the fiscal system.

The Leviathan model does, however, have some straightforward implications for observable fiscal behavior. It is the purpose of this paper to examine one of these testable implications. Brennan and Buchanan stress that fiscal decentralization is itself a powerful constraint on Leviathan: competition among governments in the context of the "interjurisdictional mobility of persons in pursuit of 'fiscal gains' can offer partial or possibly complete substitutes for explicit fiscal constraints on the taxing power" (1980, p. 184). Such competition among governments in a federal system that places heavy reliance on "local" fiscal decisions will greatly limit the capacity of Leviathan to channel resources into the public sector. In short, as indicated by the epigraph to this paper, the Leviathan model implies that, other things equal, the size of the public sector should vary inversely with the extent of fiscal decentralization.[2]

*Department of Economics and Bureau of Business and Economic Research, University of Maryland, College Park, MD 20742. I am grateful to Mark Eiswerth and Christopher Graves for their excellent assistance with the computations; to Jonathan Levin of the Government Finance Statistics Division of the International Monetary Fund for his invaluable help with the data; and to the Alfred P. Sloan Foundation for its support of this work. For helpful comments on an earlier draft, I thank Fred Abraham, Charles Brown, Bruce Hamilton, Harry Kelejian, Michael Luger, Edwin Mills, Richard Musgrave, Daniel Rubinfeld, Robert Schwab, and participants in the George Mason Public Choice Seminar and the Sloan Workshop in Urban Public Economics at the University of Maryland.

[1]For a critical appraisal of the Leviathan model, see Richard Musgrave (1981).

[2]As Musgrave has pointed out to me, other sorts of models besides Leviathan could produce such an outcome. He notes, in particular, the redistribution function. Under a highly decentralized public sector, there is likely to be comparatively little in the way of assistance to the poor for two reasons. First, sorting out along Tiebout lines will imply relatively income-homogeneous jurisdictions with little scope for redistribution from wealthy to poor *within* jurisdictions. And, second, the fear of attracting mobile poor with relatively generous support programs will tend to deter the adoption of such programs. All this suggests that the scope for public relief programs will be more circumscribed under a relatively decentralized fiscal system. This would lead us to expect comparatively larger budgets where the public sector is more centralized as a result of a greater demand for assistance to low-income households.

I. More on the Empirical Test

Brennan and Buchanan thus see a decentralized public sector as a mechanism for limiting the growth or size of government. But there is an alternative view. Suppose that instead of a monopolistic setting, competition among political parties produces an outcome that conforms fairly closely to the tastes of the citizenry—as under the conventional median-voter model. In such a competitive political environment, one would have no reason to expect a negative association between the size of public budgets and the degree of fiscal decentralization. In a centralized setting, the outcome would conform to the preferences of the "overall" median voter. If, in contrast, levels of output were set independently in each jurisdiction, the median voter in each locality would effectively choose the budget. It is impossible to determine whether the average level of output in the decentralized case would exceed or fall short of output under centralized decision making without knowing both the distribution of tastes and the location of the populace.[3]

In fact, one might argue for the competitive case that, *from a purely budgetary perspective*, increased fiscal decentralization would typically result in a higher level of government expenditure. Greater decentralization may result in the loss of certain "economies of scale" with a consequent increase in costs of administration. This, of course, need not imply an inferior outcome; the welfare gains from the tailoring of local budgets to local preferences and from the wider range of choice available to mobile consumers may more than offset the additional administrative expenditures. The point is simply that fiscal decentralization may be relatively expensive in budgetary terms. Thus, a more competitive view of the functioning of the public sector would suggest, contrary to the Leviathan model, an absence of a positive association between government size and fiscal centralization with the possibility that this association might even be negative.

John Wallis, an American economic historian, has suggested to me an even stronger hypothesis. He contends that since individuals have more control over public decisions at the local than at the state or national level, they will wish to empower the public sector with a wider range of functions and responsibility where these activities are carried out at more localized levels of government. Based on his reading of American history, Wallis offers the conjecture that over time and across states, the state-local sector has tended to be larger, the more decentralized is fiscal decision making.

The resolution of this debate requires recourse to the actual facts of governmental structure and budgetary outcomes. A little over a decade ago in the context of a larger study. I undertook a cursory examination of the relationship between the size of government and the extent of fiscal decentralization, using a cross-sectional sample of 57 countries (1972, pp. 209–13). Regressing a measure of the size of the public sector (tax revenues as a fraction of national income) on a fiscal centralization ratio (i.e., central-government tax revenues as a percentage of total tax revenues), I found a strong and statistically significant negative association. Increased fiscal decentralization in this simple regression equation was associated with a *larger* government sector. However, after controlling for the effect of the level of income on the size of government (i.e., Wagner's "Law"), the relationship between the two variables of interest became much weaker: the sign remained negative but was not significantly different from zero at the usual confidence levels. This earlier work thus

[3]As Musgrave suggested to me, the result is likely to depend critically on the arrangement of preferences in the tails of the distribution. Suppose, for example, that the upper tail of the distribution of tastes stretches out quite far. This will have little impact on the outcome under centralized decision making, since the distribution of the tail has no effect on the median voter. Under a more decentralized system, however, these high demanders might reside together in a jurisdiction with an extraordinarily high level of output. Musgrave conjectures that the tail on the high side might be expected to be the more skewed so that, *ceteris paribus*, the budget is likely to be larger under decentralization.

does not support the Brennan-Buchanan model of Leviathan; in none of the statistical tests did the results suggest that fiscal decentralization was significantly associated with a relatively small public sector.

In view of the renewed interest in this issue, I turn in this paper to a more careful study of this relationship, making use of two quite different bodies of data. First, the International Monetary Fund (1982) has recently provided an extraordinarily rich set of information on public finances disaggregated by level of government. From this data, I have been able to assemble measures of the extent of fiscal centralization for a sample of 43 countries. The second sample is quite different: it consists of the state-local sector in each of the 48 contiguous states in the United States. For this second sample, I explore the association between the budgetary size of state and local government in each state and the degree of decentralization of the state-local "fisc."

Each of the two samples has its relative strengths and weaknesses for purposes of this study. The international sample encompasses much greater diversity in governmental structure and consequently provides considerably more variation in the variables of interest. There is, in a sense, more to work with here. However, this comes at some cost, for the state-local data are undoubtedly more reliable and comparable than those from different countries; there should, therefore, be less in the way of measurement error and differing classifications of budgetary items. Finally, there is the issue of the extent of fiscal mobility. There must typically exist greater mobility across state than national boundaries suggesting that fiscal decentralization, from the Brennan-Buchanan perspective, should enforce a greater fiscal "discipline" on state than on national governments. In short, the scope for state governments to extract "surplus" from their residents is probably less than for the central government. Nevertheless, state governments are surely in a much less constrained position than are the myriad of smaller local governments that compete with one another within a state's borders. The Leviathan model would thus predict that, other things equal, those states with a more decentralized fiscal structure should have a smaller state-local sector.

The testing procedure will be to take as the null hypothesis the proposition that government size and the extent of decentralization bear no relation to one another. Since the various views discussed above suggest that the relationship between these two variables could be either negative or positive, it is appropriate to use a two-tailed test to determine if we can reject the null hypothesis in favor of the alternatives.

II. Empirical Results: The State-Local Sector

As a measure of the "size" of the public sector, the Leviathan view suggests that we focus on the level of tax revenue that the state extracts from the economy. Normalizing for the level of income, I take as the dependent variable for this part of the study aggregate state-local tax receipts in each state as a fraction of personal income (G). The appropriate measure of the extent of decentralization is less clear. In consequence, I have used three plausible indices of state-local decentralization. The first two are fiscal centralization ratios: the state share of state-local general revenues (R, a revenue measure) and the state share of state-local total expenditure (E, an expenditure measure). As a third and a nonfiscal index of decentralization, I have employed the absolute number of local government units in the state (L).[4] This variable is suggested by Brennan and Buchanan's observation that "the potential for fiscal exploitation varies inversely with the number of competing governmental units in the inclusive territory" (1980, p. 185).[5]

[4] I experimented with some variants of L involving the normalization of the number of local governments for land area and for population size. In these forms, its explanatory power was considerably reduced compared to its unnormalized form.

[5] In an interesting theoretical paper, Dennis Epple and Allan Zelenitz (1981) have explored the extent to which competition among local jurisdictions can limit the power of local governments to extract tax revenues from their residents. They find that increasing the number of local jurisdictions limits the scope for such taxation, but cannot eliminate it entirely.

TABLE 1—SUMMARY STATISTICS

Variable	Mean	Maximum	Minimum	*SD*
G	.12	.18	.10	.02
R	.58	.78	.43	.08
E	.43	.59	.22	.08
L	1660	6620	120	1450

Note: G = total state-local tax receipts as a fraction of state personal income; R = state share of state-local general revenues; E = state share of state-local total expenditures; L = number of local government units; and SD = standard deviation.

Table 1 reports the basic summary statistics for the four variables. State-local tax revenues as a fraction of personal income (G) have a mean value of .12 and vary from a high of .18 for New York to a low of .10 for Ohio. There is considerable diversity in the extent of fiscal concentration: the centralization ratio for revenues (R), for example, varies from a maximum of .78 for New Mexico to a minimum of .43 for New Hampshire with a mean value of .58. Likewise, the number of local governments ranges from 6,620 in Illinois to only 120 in Rhode Island. The sources of the data are listed in the Appendix. I would note that the fiscal data come from the *1977 Census of Governments*. It seemed advisable to use a year prior to the "disturbances" introduced by the numerous measures enacted under the fiscal-limitation movement in the United States.

As a first and admittedly crude examination of the associations among these variables, I looked at the simple rank correlation between G and each of the three measures of state-local decentralization. While this fails to hold constant for the influence of other variables, it does give some sense of the simple relationship between the pairs of variables without imposing any a priori assumptions concerning the probability distribution of the population from which the sample was drawn. Table 2 reports the value of the Spearman rank correlation coefficient (Spearman's *rho*) between G and each of the decentralization measures and the associated t-statistic (Sidney Siegel, 1956, p. 202). The correlation between G and each of the fiscal centralization ratios is negative, indicating that a more centralized state-local sector

TABLE 2—SPEARMAN RANK CORRELATION COEFFICIENT

Variable Pair	Correlation Coefficient	t-Statistic
G, R	−.22	1.50
G, E	−.25	1.73
G, L	−.06	0.41

tends to be associated with a *smaller* state-local sector. Although the negative correlation runs counter to the prediction of the Leviathan model, the association is not sufficiently strong in either case to reject the null hypothesis of a zero correlation at a .05 significance level using a two-tailed test.[6] In contrast, the correlation between G and L, the number of local governments, has the negative sign implied by the Leviathan hypothesis; however, the relationship is very weak, and again we cannot reject the null hypothesis of a zero correlation. The simple rank correlations thus do not support the Leviathan model.

In an attempt to control for other variables that influence the size of the public sector, I move next to a multiple-regression analysis. In the absence of a fully specified model of the economy, I have resorted to a series of admittedly *ad hoc*, reduced-form equations using explanatory variables that other studies have found to be of significance in explaining the size of the government sector.[7] The regression analysis requires one further modification: since the dependent variable has a range limited to the zero-to-one interval, the basic assumptions of the regres-

[6]As an alternative testing procedure, I might have focused on the Brennan-Buchanan prediction of a positive relationship between G and the measure of fiscal centralization and taken it as the alternative to the null hypothesis of no relationship. This procedure would imply a one-tailed test of H_0 against H_1. However, as the regression results will make clear, this form of test would not alter the basic findings.

[7]Although I am unhappy with my inability to derive formally the equation to be estimated from a fully specified structural model, I would note that the dependent variable is the ratio of government revenues to personal income. The complete structural model would thus have to determine both the level of government budgetary activity and the level of private income—a formidable task.

TABLE 3—ESTIMATED REGRESSION EQUATIONS

(1*R*)	$G' = -1.8 - .004R$	$R^2 = .04$
	(11.5) (1.5)	
(1*E*)	$G' = -1.7 - .006E^*$	$R^2 = .10$
	(14.9) (2.2)	
(1*L*)	$G' = -2.0 + .5 \times 10^{-5}L$	$R^2 = .003$
	(59.2) (0.3)	
(2*R*)	$G' = -2.9 + .0001Y^* - .003P - .002U + .01I^* - .006R$	$R^2 = .32$
	(8.7) (3.5) (0.4) (1.7) (2.8) (1.7)	
(2*E*)	$G' = -2.8 + .0001Y^* - .0002P - .002U + .008I - .004E$	$R^2 = .29$
	(6.6) (3.1) (0.0) (2.0) (1.9) (1.0)	
(2*L*)	$G' = -3.0 + .0001Y^* - .008P - .002U^* + .007I - .00002L$	$R^2 = .29$
	(9.2) (3.7) (1.0) (2.1) (1.8) (1.1)	

Note: The numbers in parentheses below the estimated coefficients are the absolute values of the t-statistic. An asterisk indicates that the estimated coefficient is statistically significant at the .05 level (using a two-tail test). U = percentage of state's population residing within Standard Metropolitan Statistical Areas (SMSA); P = population (in millions); Y = state personal income per capita; I = intergovernmental grants as a percentage of state-local general revenues.

sion model are not satisfied. To correct for this, I have used the logistic transformation to create a new dependent variable whose value can range over the whole set of real numbers (see, for example, Robert Pindyck and Daniel Rubinfeld, 1981, p. 287). Table 3 reports the findings for the transformed G (denoted G'). The results, incidentally, do not differ substantively from those if the transformation is not used.

The first three equations in Table 3 are the simple regressions of G' on each of the measures of decentralization. They are roughly consistent with the rank correlations, indicating negative associations of G' with R and E. The simple correlation with L is now positive but remains very weak. The next three equations attempt to control for the influence of other key variables on the size of the public sector. The level of per capita income (Y), for example, has a positive and significant association with the size of the public sector (consistent with earlier studies of Wagner's Law). Population size (P) exhibits a postivie sign but is not significantly different from zero in any of the three equations. Next, the extent of urbanization (U) is negatively related to G' and is statistically significant in one of the three equations. Other things equal, the more urbanized a state, the smaller is its public sector, reflecting perhaps some economies in providing services to more densely populated areas.[8] There exists a body of theoretical and econometric work suggesting that intergovernmental grants provide a significant stimulus to expenditures by the recipient. The findings here are consistent with this. The variable I, the percentage of state-local general revenues that comes from intergovernmental grants, has the expected positive coefficient and is statistically significant in equation (2R).

Of central interest here is the effect of including these control variables on the measured influence of the decentralization variables. The fiscal centralization ratios, R and E, retain their negative sign, but in neither case can we reject the null hypothesis of no association. The number of local governments L now has a negative sign (consistent with the Leviathan view) and a somewhat larger t-statistic, but we are likewise unable to reject the null hypothesis of no association at the .05 level. The results of the multiple-regression analysis do not appear to provide

[8]Rubinfeld has suggested that the negative and significant sign of the estimated coefficient of the urbanization variable could be interpreted as providing support for Leviathan. The contention is that in more highly urbanized areas, the Tiebout process of sorting can better work itself out. Thus, more highly urbanized states are likely, from this perspective, to be effectively more decentralized.

real support for either the view that decentralization constrains the size of the public sector or that it results in a more expansive government sector. I would note, moreover, that these findings seem quite robust. I experimented with several different specifications of the multiple-regression equation, including the use of some nonlinear transformations of the key variables. In a few instances, the fiscal centralization variables were both negative and statistically significant, providing some support for the Wallis hypothesis. But in no cases did I find any significant coefficients with the sign predicted by the Leviathan model.[9]

III. Empirical Results: The World Sample

The procedures for analyzing the data from my international sample of 43 countries are essentially the same as those for the state-local study. First I have computed the Spearman rank correlation coefficients between a revenue measure of government size and my measures of fiscal centralization. And second, I present estimates of regression equations using basically the same control variables as earlier. In addition to results for the world sample as a whole, I report the estimated equations for two subsamples: a group of 18 industrialized countries (so classified by the IMF) and the remaining 25 "developing" countries. As will become apparent shortly, there are some striking and important differences between the developing and industrialized countries in terms of both the size and structure of the public sector.

TABLE 4—SUMMARY STATISTICS

Variable	Mean	Maximum	Minimum	*SD*
(a) World Sample ($N = 43$)				
G	.31	.57	.12	13
R	.85	.99	.48	14
E	.79	.99	.40	18
(b) Industrial Countries ($N = 18$)				
G	.42	.57	.26	.09
R	.76	.96	.48	15
E	.65	.92	.40	16
(c) Developing Countries ($N = 25$)				
G	.22	.50	.12	.09
R	.92	.99	.71	08
E	.89	.99	.63	11

Note: G = total public revenues as a fraction of GDP; R = central government share of total government revenue; E = central government share of total public expenditure.

Table 4 presents the summary statistics for the world sample. The government-size variable G (here defined as total public revenues divided by gross domestic product) exhibits considerably more variation than in the set of state-local data. Tax revenues as a fraction of GDP have a mean value of .31 and range from a high of .57 in Sweden to a low of .12 in Bangledesh. For the international sample, I have only fiscal measures of the extent of centralization: R is the fraction of total general revenues going to the central government, and E is the fraction of total public expenditure attributable to the central government. For purposes of E, intergovernmental grants are excluded from the grantor's expenditures so that E is the central government's share of total disbursements. As Table 4 indicates, the extent of fiscal centralization ranges widely within the sample, reaching from almost complete centralization of .99 to a central government share of well under one-half.

Panels (b) and (c) of Table 4 reveal the dramatic differences between the typical industrialized and developing country. The industrialized countries exhibit both a much larger size of the public sector (a mean value of G of .42 compared to .22) and a far less centralized government sector (a mean of E

[9] There is a further issue. Since some grant funds take a matching form, the level of public spending may influence the amount of intergovernmental grants-in-aid. Intergovernmental grants (I) thus may be taken to be an endogenous variable in the multiple-regression equations implying the presence of simultaneous-equation bias. Likewise, one can make a case for the endogeneity of the fiscal centralization ratios, R and E. To address this matter, I reestimated equations ($2R$), ($2E$), and ($2L$) using the two-stage least squares procedure ($2SLS$) and treating I, R, and E as endogenous variables. In the $2SLS$ equations, the estimated coefficient for the grant variable remained positive but its t-statistic declined somewhat. The estimated coefficients of the centralization variables (R, E, and L) retained their negative signs but were not significantly different from zero.

of .65 compared to .89). The developing countries are characterized by relatively small, but highly centralized, public sectors.

A listing of the data and their sources appears in the Appendix. I have used IMF data to construct the fiscal centralization ratios. From the IMF *Government Finance Statistics Yearbook* (1982), I included in this study every country for which I had reasonable assurance that the data encompassed all the relevant levels and units of government. I computed the centralization ratios for the most recent year for which data were available—1980 or 1981 in most cases—but reaching back as far as 1976 in a few instances. This produced the sample of 43 countries. Regarding the nonfiscal data, I have used as a measure of per capita income (in U.S. dollars) the recent estimates by Robert Summers et al. (1980) from the International Comparison Project of the United Nations.

Table 5 reports the rank correlation coefficients and associated t-statistics between G and R and between G and E for the entire world sample and for the two subsamples. For the entire set of 43 countries, the rank correlation between the size of the public sector and the extent of centralization is strongly and significantly negative. A relatively decentralized public sector is typically comparatively large. However, this result is misleading. It reflects the fact just noted that the poorer developing countries have small, centralized public sectors, while the industrialized countries have relatively large and decentralized governments. When we examine the coefficients for the two subsamples, we find that there is no longer a significant relationship between the variables. Within the subsamples, government size seems to have little relation to the degree of centralization in the public sector.

Turning next to the regression analysis in Table 6 (where again I make use of the logistic transformation of G, denoted G'), note first that the simple regressions ($1R$) and ($1E$) confirm the rank-correlation analysis. For the whole sample, G is negatively and significantly related to both R and E, but within each subsample, this association effectively disappears. Equations ($2R$) and

TABLE 5—SPEARMAN RANK CORRELATION COEFFICIENT

Variable	Correlation Coefficient	t-Statistic
(a) World Sample		
G, R	$-.39$	2.71
G, E	$-.49$	3.60
(b) Industrialized Countries		
G, R	$-.02$	0.08
G, E	$-.15$	0.61
(c) Developing Countries		
G, R	.20	0.98
G, E	.12	0.58

($2E$) are the multiple-regression equations, where the control variables are basically the same as in the state-local equations. Within each of the samples, equation ($2R$) is virtually identical with ($2E$), suggesting that it makes little difference whether we use a revenue or expenditure measure of the extent of fiscal centralization. For the entire world sample, the multiple-regression equations have substantial explanatory power: they can "explain" nearly 80 percent of the variation in the size of the public sector. In particular, I find that other things equal, high-income countries typically have relatively large public sectors, populous nations tend to have comparatively small government sectors, and countries that rely heavily on intergovernmental grants have, on average, large public sectors. However, the measures of fiscal centralization, R and E, contribute virtually nothing to the explanatory power of the equations. The extent of centralization in the public sector appears to have little effect on the size of government.[10]

[10] Responding to the empirical findings for the international sample, Edwin Mills has offered the interesting suggestion that to measure Leviathan's monopoly power, the fiscal autonomy of decentralized governments is probably a better measure than their share of taxes or spending. Mills suggests a constitutional variable: the existence of a federal constitution. Following up on this, I find that if I divide the total sample into 8 federal and 35 nonfederal countries, the mean size of the public sector as a fraction of GDP (i.e., G in the earlier equations) is slightly larger for federal countries. Probably of more relevance, however, within the subsample of 18 industrialized countries, the mean value of G is somewhat less for the 6 federal countries than for the 12

TABLE 6—ESTIMATED REGRESSION EQUATIONS

(a) World Sample								
(1R)	$G' =$	0.8	$-2.0R^*$					$R^2 = .16$
		(1.3)	(2.8)					
(1E)	$G' =$	0.6	$-1.9E^*$					$R^2 = .26$
		(1.6)	(3.8)					
(2R)	$G' =$	-1.9	$+.0003Y^*$	$-.004P^*$	$+.001U$	$+0.6I^*$	$+0.1R$	$R^2 = .78$
		(3.7)	(5.0)	(2.8)	(0.4)	(2.5)	(0.2)	
(2E)	$G' =$	-1.9	$+.0003Y^*$	$-.004P^*$	$+.001U$	$+0.6I^*$	$+0.1E$	$R^2 = .78$
		(4.6)	(5.2)	(2.9)	(0.4)	(2.7)	(0.3)	
(b) Industrialized Countries								
(1R)	$G' =$	-0.3	$-.07R$					$R^2 = .001$
		(0.5)	(0.1)					
(1E)	$G' =$	$-.003$	$-0.5E$					$R^2 = .04$
		(0.0)	(0.8)					
(2R)	$G' =$	-1.8	$+.0004Y^*$	$-.005P^*$	$-.007U$	$+0.7I$	$+0.4R$	$R^2 = .57$
		(2.3)	(3.0)	(3.1)	(1.2)	(1.8)	(0.5)	
(2E)	$G' =$	-1.9	$+.0004Y^*$	$-.005P^*$	$-.007U$	$+0.8I^*$	$+0.4E$	$R^2 = .57$
		(2.6)	(3.2)	(3.2)	(1.3)	(2.3)	(0.7)	
(c) Developing Countries								
(1R)	$G' =$	-2.2	$+0.9R$					$R^2 = .02$
		(1.8)	(0.7)					
(1E)	$G' =$	-1.5	$+0.3E$					$R^2 = .004$
		(1.9)	(0.3)					
(2R)	$G' =$	-2.9	$+.0004Y$	$+.0003P$	$+.004U$	$+0.4I$	$+1.0R$	$R^2 = .59$
		(2.2)	(1.6)	(0.1)	(0.5)	(1.2)	(0.7)	
(2E)	$G' =$	-2.7	$+.0004Y$	$+.0002P$	$+.004U$	$+0.5I$	$+0.7E$	$R^2 = .59$
		(2.9)	(1.6)	(0.1)	(0.5)	(1.6)	(0.8)	

Note: The numbers in parentheses below the estimated coefficients are the absolute values of the associated t-statistic. An asterisk indicates that the estimated coefficient is statistically significant at the .05 level (using a two-tail test); Y = income per capita in U.S. dollars for 1977; P = population (in millions); U = percentage of population living in urban areas; I = intergovernmental grants as a percentage of total government general revenues.

nonfederal nations—a result that is presumably consistent with the Leviathan hypothesis. However, in neither case could a simple test for the difference between the means reject the null hypothesis that the observations come from the same population (at a .95 level of confidence). I also reestimated the regression equations in Table 6 substituting for the fiscal centralization ratios a dummy variable equal to one for those countries with a federal constitution and zero otherwise. In the multiple regression equations, the estimated coefficients on the dummy variable were of negative sign. Taking the point estimate for this coefficient from the equation for the industrialized countries, one finds that, *ceteris paribus*, the value of G for a federal country is about 5 percentage points less than for a nonfederal nation. However, in none of the equations was the t-statistic of sufficient size to reject the null hypothesis of no association. Related to this, Richard Bird (1984) using a sample of 13 countries from a study by Morris Beck (1981) finds that, over the period 1950–77, the rate of growth of real government expenditure in relation to GNP was actually slightly higher in federal than in nonfederal countries. It seems hard to find any really significant differences here.

The multiple-regression equations for the two subsamples tell pretty much the same story (except that the t-ratios for the subsample of developing countries tend to be somewhat smaller). The fiscal-centralization variables have positive coefficients, but the t-ratios remain quite small. Finally, like the results for the state-local analysis, the findings for the international sample seem quite robust. The use of a number of alternative specifications of the equations and variables did not alter the substance of the results.[11]

[11]As with the state-local sample, I reestimated the multiple-regression equations using $2SLS$ and treating intergovernmental grants (I) and the centralization ratios (R and E) as endogenous variables. The general pattern of the results remained similar to those reported in Table 6 although there was a substantial reduction in the t-statistic for the grant variable (I). The estimated coefficients for the centralization ratios remained statistically insignificant.

IV. Conclusion

Overall, the results of this study suggest that there does not exist a strong, systematic relationship between the size of government and the degree of centralization of the public sector. At the offset, I set forth as the null hypothesis the proposition that centralization and the size of government have little to do with one another. We certainly cannot reject this proposition from the findings in any of the samples or subsamples in this study.

What implications can we draw from all this? I would stress that the basic finding does not imply that there is no place in public economics for the revenue-maximization hypothesis. There is, in fact, considerable evidence to support budget-maximizing behavior by public agents in certain sorts of institutional settings (for example, the literature on bureaucracy). But it is another matter to try to characterize the entire public sector as a monolithic, monopolistic actor with the sole objective of making the government sector as large as possible. The results, it seems to me, cast considerable doubt on the usefulness of the Leviathan model. If, in fact, potentially pervasive revenue-maximizing forces are at work in the public sector, we should expect to see these forces manifest themselves in terms of larger budgets where given the opportunity to do so. As Brennan and Buchanan suggest, Leviathan will have much more scope for action in a relatively centralized public sector. But I seem to find no real difference in outcomes whether Leviathan is constrained by decentralization or not. Perhaps, after all, Leviathan is a mythical beast.

Appendix

Sample of countries:

Australia
Austria
Bangledesh
Brazil
Canada
Chile
Costa Rica
Cyprus
Denmark
Ecuador
Ethiopia
Fiji
Finland
France
Germany (West)
Greece
Honduras
Iceland
Iran
Ireland
Israel
Kenya
Korea
Luxembourg
Malawi
Mauritius
Mexico
Netherlands
New Zealand
Norway
Pakistan
Panama
Paraguay
Philippines
South Africa
Spain
Sweden
Switzerland
Thailand
Tunisia
United Kingdom
United States
Uruguay

Sources of data for the state-local study:

Fiscal data: *1977 Census of Governments*, Vol. 4, No. 5, Tables 23, 24, 35, and 39. The same source was used for total population, urban population, number of local governments, and income per capita, but came from Table 46. The land area, population density, and date entered the union is from *The World Almanac...*, 1980.

Sources of data for the international study:

Fiscal data: IMF *Government Statistics Yearbook*, Vol. VI, 1982; the income per capita: Summers et al., pp. 19–66 (data used for 1977); *GDP*: United Nations *Yearbook of National Accounts Statistics, 1980*, Vol. 1, 1982; total population: *United Nations Demographic Yearbook 1981*, Table 6, 1983; urban population was from the same source but supplemented by *The World Almanac...*, 1983.

REFERENCES

Beck, Morris, *Government Spending, Trends and Issues*, New York: Praeger, 1981.

Bird, Richard M., "Federal Finance in Comparative Perspective," Working Paper No. 84–22, Department of Economics and Institute for Policy Analysis, University of Toronto, June 1984.

Brennan, Geoffrey, and Buchanan, James, *The Power to Tax: Analytical Foundations of a Fiscal Constitution*, Cambridge; New York: Cambridge University Press, 1980.

______ and ______, "Tax Instruments as Constraints on the Disposition of Public Revenues," *Journal of Public Economics*, June 1978, *9*, 301–18.

______ and ______, "Towards a Tax Constitution for Leviathan," *Journal of Public Economics*, December 1977, *8*, 255–73.

Epple, Dennis and Zelenitz, Allan, "The Implications of Competition Among Jurisdictions: Does Tiebout Need Politics?," *Journal of Political Economy*, December 1981, *89*, 1197–1217.

Musgrave, Richard, "Leviathan Cometh—or Does He?," in H. Ladd and T. N. Tideman, eds., *Tax and Expenditure Limitations*, COUPE Papers on Public Economics, *5*, Washington: The Urban Institute, 1981, 77–120.

Oates, Wallace, *Fiscal Federalism*, New York: Harcourt Brace Jovanovich, 1972.

Pindyck, Robert and Rubinfeld, Daniel, *Econometric Models and Economic Forecasts*, 2d ed., New York: McGraw-Hill, 1981.

Siegel, Sidney, *Nonparametric Statistics for the Behavioral Sciences*, New York: McGraw-Hill, 1956.

Summers, Robert, Kravis, Irving and Heston, Alan, "International Comparison of Real Product and its Composition: 1950–77," *Review of Income and Wealth*, March 1980, Series 26, No. 1, 19–66.

International Monetary Fund, *Government Finance Statistics Yearbook*, Vol. VI, Washington: International Monetary Fund, 1982.

United Nations, *Yearbook of National Accounts Statistics, 1980*, Vol. 1, New York: United Nations, 1982.

______, *Demographic Yearbook, 1981*, New York: United Nations, 1983.

U.S. Bureau of the Census, *1977 Census of Governments*, Vol. 4, No. 5: *Compendium of Government Finances*, Washington: USGPO, August 1979.

The World Almanac and Book of Facts, 1980; 1983.

Industry Structure with Fluctuating Demand

By David E. Mills and Laurence Schumann*

The intra-industry distribution of demand-induced output fluctuations is a neglected aspect of the theory of industrial organization. The standard model of competitive equilibrium has identical firms and stationary demand. Eytan Sheshinski and Jacques Drèze (1976) extended this model to an environment with demand fluctuations to study its industry-structure implications. They concluded that demand fluctuations increase the number of firms sustained in competitive equilibrium, and that each firm produces less on average with fluctuations than without. By preserving the identical-firm assumption, Sheshinski and Drèze predict that industry-wide output fluctuations are distributed equally among competitors.

An alternative model of competitive equilibrium with stationary demand would have constant-returns-to-scale firms with various output levels. Such firms would be "proportionate clones" of each other in a technological sense. The analagous extension of this model to an environment with demand fluctuations would have output fluctuations distributed equiproportionately among competitors; a firm whose mean output is 4 percent of the industry mean would preserve its market share and absorb 4 percent of industrywide fluctuations in output. Neither of these theories—that fluctuations are distributed equally or equiproportionately—are totally convincing. (This is especially so in light of evidence presented below.) In this paper we propose a more complete and slightly more subtle theory of industry structure with demand fluctuations, one that gives a larger place to the technological diversity of rivals.

We argue that a significant part of the general pattern of firm heterogeneity observed in industries is due to accommodating demand fluctuations.[1] Further, these differences are perfectly consistent with a competitive theory of industry structure where rivals' strategies are seen to include the choice of a production technology. A basic assumption of the paper is that the space of available technologies affords a tradeoff between static-efficiency and *flexibility*, a characteristic of firms first described by George Stigler (1939) that enhances their ability to vary output.[2] With stationary demand and a constant price, all firms in a competitive industry share a common minimum level of average costs. Only static-efficient rivals survive in the long run. But, with fluctuating demand, we show that it is possible for firms with higher minimum average cost also to survive if they are sufficiently flexible. This tradeoff makes it possible for technologically diverse firms to compete with each other by relying on offsetting cost advantages.

*Department of Economics, University of Virginia, Charlottesville, VA 22901. Mills would like to thank the Earhart Foundation and the U.S. Small Business Administration for financial support. Many colleagues have provided helpful comments on earlier versions.

[1]Many intra-industry differences in firms of course owe nothing to fluctuations. Richard Caves and T. A. Pugel (1980) and Michael Porter (1979) both explore a variety of strategies that distinguish rivals. Caves and Pugel explicitly consider technological diversity but not in connection with accommodating demand fluctuations. They suggest that small firms do not use merely "less efficient miniature versions of the equipment employed by their larger rivals," but that "the long-run viability of the smaller firms is questionable unless the smaller firms possess other offsetting competitive advantages" (p. 20). Ours is a theory of one such advantage. Robert Lucas (1978) and, more recently, Walter Oi (1983) have attributed intra-industry firm-size heterogeneity to differences in the entrepreneurial ability of principals or managerial agents.

[2]There are several technological and institutional reasons for this tradeoff: cost-minimizing factor combinations for stationary output may hinder frequent and significant output adjustments; when dealing with upstream suppliers firms with stable output levels can avoid spot markets and enter into more advantageous long-term contracts, or even integrate upstream to secure steady and less costly supply; alternatively, they may forgo maintaining supply and product inventories necessitated by wide fluctuations; labor costs are likely greater for firms with greater employment variability.

Another component of the theory is that the tradeoff in the space of production technologies associates greater flexibility with smaller-sized firms. Lower minimum average costs are associated with larger-sized firms, *ceteris paribus*. Briefly, small firms rely more on variable factors than large firms, and the latter enjoy scale economies. Small rivals in fluctuation-prone industries ply a flexibility advantage and survive by absorbing a disproportionate share of industrywide output fluctuations. Conversely, large firms have output levels more stable than their industries at large.

This theory and the technological tradeoff that supports it are fully described in the next section. The main predictions of the theory are that the variability over time of sales and employment are inversely related to firm size within industries. These predictions are tested in the following section using data on a sample of 856 U.S. manufacturing firms during the period 1970–1980. We also present several sensitivity tests to control for some alternative interpretations of our findings.

I. Theory

We begin by assuming that the production technologies of firms within competitive industries are nontrivially different. This is strongly suggested by casual observation and is supported by empirical evidence. Consider for instance the finding that capital-output ratios are positively related to firm size within industries.[3] Several reasons have been offered for this finding. The most prevalent one is that large firms are more capital-intensive than small firms because they use production methods with greater capital requirements to achieve technical economies of scale. Another is that large firms usually are more vertically integrated than small ones. Other things equal, firms that are more vertically integrated have greater capital requirements since they produce intermediate products as well as final products. If size varies directly with the degree of vertical integration, then large firms would be more capital-intensive. While this is plausible theoretically, the evidence suggests that the degree of vertical integration is not the major explanatory factor.[4] There are other minor contributing factors, but none of them as compelling as the "production methods" line of reasoning, especially in view of the massive evidence that has accumulated in favor of significant technical scale economies.[5]

For an industry with diverse firms to be competitive, meaning here that all firms earn only normal returns, it follows that a pattern of offsetting cost advantages must be present. Otherwise some rivals—perhaps small firms who do not enjoy technical scale economies—would be driven from the market. One such advantage, we hypothesize, is flexibility in the face of demand fluctuations. As first described by Stigler (1939), flexibility refers to those attributes of a production technology that accommodate greater output variation. Stigler's discussion of flexibility was couched in terms of firms' cost curves, and while he did not define it formally, his notion of flexibility seems to be one that varies inversely with the curvature of total costs.[6] In the context of U-shaped average costs, flexibility is greater with flat-bottomed average costs and flat or gently inclined marginal costs. Flexibility is less when average costs rise rapidly around their minimum point and have steep marginal costs. Flexibility supplies an advantage in fluctuation-prone industries because it sustains some firms who are less static efficient than others. Specifically, a competitive equilibrium with demand fluctuations can have diverse firms coexisting with each other because of a tradeoff in the space of production technologies between flexibility and the minimum level of average costs.

To see this, suppose firms in an industry can choose a production technology from a

[3] Early evidence is surveyed by Hiram Davis (1956). Stigler (1963) and Caves-Pugel supply further evidence for U.S. manufacturing.

[4] See Matityahu Marcus (1969).

[5] See for instance J. Johnston (1960) or F. M. Scherer (1980).

[6] Thomas Marschak and Richard Nelson (1962), following Stigler (1939), define flexibility as the reciprocal of the second derivative of total costs in a model with quadratic costs.

set of discrete options $i=1,\dots,I$ each with quadratic total costs

$$(1)\qquad c_i(q_i)=\alpha_i+\beta_i q_i+q_i^2/2\gamma_i,$$

where q_i is the level of homogeneous output produced by a type-i firm.[7] The output level where average costs are minimized and the minimum level of average costs for a type-i firm are

$$(2a)\qquad q_i^*=(2\alpha_i\gamma_i)^{1/2};$$

$$(2b)\qquad c_i(q_i^*)/q_i^*=\beta_i+(2\alpha_i/\gamma_i)^{1/2}.$$

Let industrywide demand fluctuate from market period to market period, and let p be the fluctuating price that equates supply and demand. The endogenous mean and variance of p are $\bar{p}$ and σ_p^2. Suppose all firms can set their output levels in each market period to equate marginal cost to price. With this, their supply curves (for $p\geq\beta_i$) are linear:

$$(3)\qquad q_i(p)=\gamma_i(p-\beta_i).$$

In competitive equilibrium the configuration of firms must be one where all firms have zero expected profits. Let $y_i(p)$ be the profit of a type-i firm. Using equations (1) and (3) we obtain

$$(4)\qquad y_i(p)=\gamma_i(p-\beta_i)^2/2-\alpha_i.$$

To find an expression for expected profit, expand y_i around $\bar{p}$, using (4):

$$y_i(p)=\gamma_i(\bar{p}-\beta_i)^2/2-\alpha_i$$
$$+\gamma_i(\bar{p}-\beta_i)(p-\bar{p})+\gamma_i(p-\bar{p})^2/2,$$

and take its expectation

$$(5)\qquad E[y_i(p)]=\gamma_i(\bar{p}-\beta_i)^2/2-\alpha_i+\gamma_i\sigma_p^2/2.$$

For all types of firms included in an equilibrium configuration, $E[y_i(p)]$ must be zero. Thus for all included firm types i, equation (5) implies that

$$(6)\qquad (\bar{p}-\beta_i)^2+\sigma_p^2=2\alpha_i/\gamma_i.$$

Combining equation (6) with (2b), we get the relation

$$(7)\qquad c_i(q_i^*)/q_i^*=\left[(\bar{p}-\beta_i)^2+\sigma_p^2\right]^{1/2}+\beta_i$$

which also must hold for all firm types i represented in an equilibrium configuration.

The degree of flexibility of a type-i firm can be represented by its elasticity of supply η_i at the price $\bar{p}$. This measure captures Stigler's notion that flexibility should be inversely related to the slope of firms' marginal cost curves. Using $\eta_i=q_i'(\bar{p})\bar{p}/q_i(\bar{p})$ and equation (3) we have

$$(8)\qquad \eta_i=\bar{p}/(\bar{p}-\beta_i).$$

(Where the denominator must be positive for firm types that produce any output at all.) The tradeoff between flexibility and minimum average costs is established by noting from (7) and (8) that

$$(9a)\qquad d\left(c_i(q_i^*)/q_i^*\right)/d\beta_i>0;$$

$$(9b)\qquad d\eta_i/d\beta_i>0.$$

If competitive equilibrium supports technologically heterogenous firms, that is, firms with different cost parameters, then more flexible firms have greater minimum average costs.[8]

Flexibility also is related to the capital intensity of firms' production technologies.

[7] If total costs are not quadratic, and unless the price distribution is normal, equations (3)–(8) and (12) are only second-order approximations. This does no violence to the tradeoffs described here.

[8] What this establishes is that if technologically heterogeneous firms coexist in a competitive fluctuation-prone industry, they exhibit a tradeoff between flexibility and static efficiency. A reasonable question to ask is whether it is possible even with fluctuating demand that firms are homogeneous. Mills (1984) has shown that the answer would be yes if the tradeoff between flexibility and static efficiency were continuous in the space of available technologies. But with a discrete set of options, as we have assumed here, continuity is supplied by linear combinations of the discrete options; thus heterogeneous firms can be expected to coexist.

Greater flexibility is achieved by increased reliance on variable factors of production. This would imply, ordinarily, that flexible technologies have lower capital requirements than inflexible technologies since capital is intrinsically less variable than labor. In terms of costs, greater reliance on capital increases fixed costs, so on a per-unit-of-output basis, fixed costs should be greater for inflexible technologies. This too can be demonstrated for the quadratic cost case. Equation (6) implies that $d\beta_i/d(\alpha_i/\gamma_i)<0$ holds among firm types i included in an equilibrium configuration. Together with inequality (9b) this implies that

$$(10) \qquad d\eta_i/d(\alpha_i/\gamma_i)<0,$$

for all i included. Also, from (2a) fixed costs per-unit-of-output, a correlative of capital intensity, is

$$\alpha_i/q_i^* = (\alpha_i/2\gamma_i)^{1/2}$$

evaluated at $q=q_i^*$. From this we have

$$(11) \qquad d(\alpha_i/q_i^*)/d(\alpha_i/\gamma_i)>0.$$

Inequalities (10) and (11) together establish that fixed costs per-unit-of-output are inversely related to flexibility among firms with heterogenous cost structures. The firm types supported by competitive equilibrium exhibit a tradeoff between flexibility and capital intensity.

The other major part of our hypothesis is that, within industries, flexibility varies inversely with firm size. The reason for this is that minimum average costs most likely are lower for large firms because of technical scale economies. It follows that if large firms are more static efficient than small ones, the comparative advantage of small rivals, given the tradeoff above, is flexibility. This hypothesis goes hand in hand with the empirical finding mentioned above that capital-output ratios are greater for large firms within industries. We have just seen that if diverse firms are included in a competitive equilibrium, the more flexible firms are less capital intensive. Large firms, we suggest, utilize relatively more capital, and more specialized forms of capital, than small competitors. This curtails their flexibility by reducing their dependence on variable factors like labor, and increasing their dependence on equipment and procedures with low adaptive capability. Small firms retain greater flexibility by relying more on variable factors.[9]

The hypothesis that small firms choose more flexible production technologies while their large counterparts choose more capital-intensive and static-efficient technologies has several empirical implications. One of them is that firm size and rate-of-return variability vary inversely within industries. This finding has been detected repeatedly in cross-sectional and longitudinal studies of the U.S. manufacturing sector and the manufacturing sector of other countries.[10] To see that this relationship follows from our hypothesis, we approximate the variance of profit for a type-i firm:

$$\sigma_{y_i}^2 \simeq [y_i'(\bar{p})]^2\sigma_p^2.$$

By Hotelling's Lemma, $y'(\bar{p})=q_i(\bar{p})$, this equation becomes

$$\sigma_{y_i}^2 \simeq [q_i(\bar{p})]^2\sigma_p^2.$$

Now let the rate of return on assets A_i for a

[9]Further support for the flexibility-size tradeoff is found in control-loss phenomena in large organizations. It is widely thought that the problem of coordinating related actions by decision makers arranged in a hierarchy becomes greater the greater are the number of decision makers and the number of levels in the hierarchy. (See Scherer or Oliver Williamson, 1967.) This means it is easier for small firms with few decision makers and simple bureaucracies to adapt quickly to changing market conditions; small firms are thus more flexible than large ones.

[10]See Sidney Alexander (1949) and J. L. McConnell (1945) for early evidence from the United States. More recent evidence is supplied by Marshall Hall and Leonard Weiss (1967), Herman Stekler (1964), Stigler (1963), and Daryl Winn (1977). Winn's study is especially interesting because he allows for an independent effect of industrial concentration on firms' variability of returns. Evidence from the U.K. is supplied by Alex Jacquemin and Wistano Saez (1976), J. M. Samuels and David Smyth (1968), and Geoffrey Whittington (1971). Jacquemin and Saez also present evidence from Japan and the European Economic Community.

type-i firm be $r_i = y_i(p)/A_i$. From this and the equation above, it follows immediately that the firm's rate-of-return variance is

$$\sigma_{r_i}^2 \simeq [q_i(\bar{p})/A_i]^2 \sigma_p^2.$$

Firms' rate-of-return variability is inversely related to the size of their capital-output ratios. Since the latter are correlated with firm size, this provides a link to explain why rate-of-return variability is inversely related to firm size.[11]

A more interesting implication of intra-industry technological diversity for our purposes has to do with how output fluctuations are distributed. From the supply curve (3) of a type-i firm we know that

$$q_i(p) - q_i(\bar{p}) = \gamma_i(p - \bar{p})$$

for any price p. This implies that the output variance of the firm is $\sigma_{q_i}^2 = \gamma_i^2 \sigma_p^2$. Multiplying both sides of this equation by $(p/q_i(p))^2$ and evaluating at $p = \bar{p}$ yields

$$cv_{q_i} = \eta_i cv_p, \tag{12}$$

where cv denotes coefficients of variation. Since η_i measures flexibility, equation (12) predicts that flexible firms' output varies more than inflexible firms' in response to fluctuating demand. If size and flexibility are inversely related as we suggest, this means that output fluctuates more, relatively, for small firms than large ones.

The proposition that size and output fluctuations vary inversely among firms within industries is tested in the next section of the paper. The test involves the relationship between firms' size and sales variability since physical output measures are not comparable across industries. Also it considers the variability of firms' sales around estimated trends of growth (or decline) to avoid picking up growth-related differences in sales variation among firms. A second implication of our hypothesis has to do with the utilization of labor, a factor more variable than capital over periods when demand and price fluctuate. If small firms are more flexible than large ones, their employment levels should be more variable. A test of whether employment variability (around trends) varies inversely with firm size within industries also is presented.

II. Evidence

Annual data on sales, assets, and employment for the period 1970–80 were obtained from Standard and Poor's COMPUSTAT industrial file for an initial sample of 1741 U.S. manufacturing firms. These firms range in size from the largest oil refiners and automobile manufacturers to enterprises with assets of less than \$1 million. COMPUSTAT assigns each firm in the file in a 2, 3, or 4-digit SIC industry code. The industries represented in the sample are not mutually exclusive; some of the 3- and 4-digit industries included are subsets of 2-digit industries also included, etc. This identifies firms' lines of business and also reflects in part the extent of firms' diversification. More diversified firms are classified in 2- or 3-digit industries while less diversified firms are classified in 4-digit industries. To insure the accuracy of firms' industry classifications, the COMPUSTAT industry codes were checked against industry classifications found in two other sources: the Securities and Exchange Commission's *Directory of Companies Required to*

[11]A number of alternative explanations have been put forward for this finding. It has been attributed repeatedly to large firms acquiring earnings stability through product diversification, where fluctuations across product lines are less than perfectly correlated. With this, diversified firms, which are larger, *ceteris paribus*, have less output variation than small specialized rivals. A second explanation is that large firms who have market power supported by scale economies use it to mitigate uncertainty. Thus they pursue strategies that reduce both the average level and variability of rates of return. Stekler, Scherer, and Caves all cite some form of this explanation. Alexander suggested that the whole pattern could be due simply to small firms being represented in greater frequency in more risky industries. This is consistent with E.A.G. Robinson's (1937) notion that volatile and uncertain market conditions limit the size of firms. However, most of the evidence gathered since Alexander finds an inverse relationship holding between firm size and rate-of-return variability within specific industries. Thus Alexander's cannot be the whole story. Winn has demonstrated that profit variability and capital intensity are negatively related across firms. Clearly this is consistent with the other findings mentioned here, the connecting link being firm size.

File Annual Reports (1980) and Dun and Bradstreet's *Million Dollar Directory* (1981). The SEC's *Directory* uses a classification scheme very close to the 3-digit SIC codes. The *Million Dollar Directory* identifies each firm with one or more (up to six) 4-digit SIC codes in descending order of importance with respect to sales. The COMPUSTAT classifications tend to be more disaggregate than the SEC's, but less so than Dun and Bradstreet's. Wherever these three sources revealed some ambiguity about the proper industry classification for a particular firm, the firm was eliminated from the sample (for example, conglomerates such as ITT which cannot be classified in any single 2-digit industry were eliminated as were firms whose COMPUSTAT classification did not appear near the top of Dun and Bradstreet's list of classifications for the firm).

Of the original 1741 firms in the sample, 650 were eliminated either because of an industry classification problem or, in the case of a few, because the COMPUSTAT file includes fewer than 3 firms with a particular industry classification. An additional 235 firms were eliminated on account of insufficient annual sales data. The final sample contains 856 firms, each identified with one of 90 industries. Eight of these are 2-digit industries and 41 each are 3- and 4-digit industries.[12]

Three variables were constructed for each firm from the 1970–80 annual observations of sales and (end of year) assets. These are the average over the period of "net plant" or fixed, tangible assets (denoted F), average annual sales (S), and firms' capital-output ratio (F/S), an indicator of average capital intensity over the decade. Another firm variable was constructed using the average value of product shipments over 1970–80 for the 90 relevant 2, 3 and 4-digit industries as taken from various issues of the U.S. Census Bureau's *Annual Survey of Manufactures*. Dividing firms' sales S by the average level of shipments in their respective industries provided a measure of their market shares (MS). Just as S and F are measures of the absolute size of firms, MS measures relative intra-industry size.

To test the hypothesis that there are systematic, firm-size effects in output variability among firms in an industry, it was necessary to compute a measure of the variability of firms' annual sales. To do this the natural logarithms of annual firm sales for 1970–80 were regressed onto a constant term and a linear time trend. Where necessary these regressions were adjusted for serial correlation. The index of firm sales variability (SV) was taken as the standard error of these adjusted regressions. Since SV measures sales fluctuation around a trend, it avoids including the effects of systematic growth or decline. Similarly, an index of industry output fluctuation (ISV) was computed by regressing the natural logarithms of industries' product shipments onto a constant and a time trend with adjustments for serial correlation.

Equation (1.1) in Table 1 shows the results of the OLS regression of the natural logarithm of SV onto a constant term and the natural logarithms of MS, F/S, and ISV.[13] (All the regressions reported in Tables 1 and 2 are OLS and use the natural logarithms of all variables.) As one might expect, firm sales variability is positively related to industry sales variability. On the other hand, firm sales variability is negatively related to capital intensity.[14] Since equation (12) indicates sales

[12]A complete list of firms and industries is available upon request from the authors.

[13]Since firms' sales and assets and industries' shipments are all reported in nominal terms, another equation was estimated based on inflation-adjusted values. Adjustments to firms' sales and industries' shipments were made using (unpublished) GNP price deflators computed and distributed by the Bureau of Economic Analysis for the relevant 2, 3, and 4-digit industries. Adjustments to firms' assets were made using the GNP price deflator for producers' durable equipment. The obvious advantage of using deflated data is that they remove that component of fluctuations due only to uneven secular price escalation. But since the deflators used are fairly aggregate estimates of the true adjustments needed for each firm and industry, using them may inject new distortions into the unadjusted data. In any case the inflation-adjusted version of the equation did not differ appreciably from the one reported.

[14]This result was repeated when F, the average "net plant" or fixed assets (end of year) over 1970–80, was replaced with the average of total assets (end of year) over 1970–80.

TABLE 1—DEPENDENT VARIABLE: SV

Equation[a]	Constant	MS	F/S	ISV	D	R^2
(1.1)	−3.459	−0.292	−0.115	0.460		0.271
[856]	(−14.95)	(−12.41)	(−1.84)	(12.61)		
(1.2)	−3.361	−0.285	−0.114	0.457	−0.062	0.272
[856]	(−12.99)	(−11.21)	(−1.82)	(12.46)	(−0.84)	
(1.3) $CR \leq 75$	−3.493	−0.299	−0.130	0.463		0.274
[830]	(−14.59)	(−12.23)	(−2.04)	(12.41)		
(1.4) $CR \leq 60$	−3.340	−0.296	−0.129	0.483		0.279
[742]	(−12.83)	(−11.45)	(−1.94)	(11.95)		
(1.5) $CR \leq 50$	−3.210	−0.327	−0.114	0.527		0.303
[629]	(−11.73)	(−11.13)	(−1.55)	(12.28)		
(1.6) $CR \leq 40$	−2.930	−0.325	−0.011	0.542		0.326
[525]	(−10.19)	(−10.38)	(−0.13)	(12.18)		
(1.7) $CR \leq 30$	−2.750	−0.297	−0.049	0.569		0.328
[385]	(−8.77)	(−8.31)	(−0.58)	(11.01)		

[a]Sample size shown in square brackets; CR = concentration ratios.

variability is greater for firms with greater flexibility, the negative relationship between SV and F/S supports our hypothesis that coexisting production technologies trade off scale economies and fixed costs for flexibility and variable costs. The most important relationship reported in equation (1.1) in Table 1 is that market share has a negative effect on sales variability. This means that larger firms within industries have less output fluctuation than their smaller rivals, and by equation (12), that flexibility and firm size are inversely related as predicted.[15,16]

While equation (1.1) in Table 1 supports the theory outlined earlier, there are two other possible explanations for the observed inverse relationship between firm size and sales variability. One of these has to do with product diversification. Unless demand fluctuations across product lines are perfectly correlated, which seems unlikely, diversified firms should experience more stable sales and output levels than specialized firms, *ceteris paribus*. Further, since large firms tend to be more diversified than small ones, an inverse relationship between size and sales variability may be attributed partially to differences in the degree of product diversification. Part of these differences—especially interindustry differences—are captured by the ISV variable since, presumably, firms grouped in aggregate 2-digit industries are more diversified than those grouped in 4-digit industries.[17] But to control more fully for

[15]Equation (1.1) in Table 1 establishes that MS has a negative effect on SV, that a firm's relative size is inversely related to sales fluctuation. Similar equations were tried using, alternatively, S and total assets as measures of *absolute* firm size. The results were not notably different than here other than diminishing the significance of the negative effect of F/S on SV. The most relevant equation is the one reported since our focus is on how fluctuations are distributed *within* industries.

[16]Along with the equations reported in Table 1 using period-average measures of sales, market shares, assets and other independent variables, similar equations were estimated using all 11 of the serial observations of these variables. Our dependent variable in these equations was the normalized absolute deviation (NAD) of firm sales:

$$NAD_{ti} = |S_{ti} - \hat{S}_{ti}|/\hat{S}_{ti}, \quad t=1,\dots 11; \quad i=1,\dots 856,$$

where the $\hat{S}$ variables are the fitted values of firm sales produced by the first set of regressions described in the text. An industry-specific version of NAD was used as an independent variable. The results of these regressions were consistent with those reported. Specifically the signs and high significance levels of all the independent variables were the same. The R^2s, however, were much lower, typically less than 0.1.

[17]To illustrate that ISV captures part of the effect of diversification, consider how ISV for each of the 8 2-digit industries included in the study compares to the weighted (by average product shipments) average of ISV for each of the 4-digit industries they encompass that also are included in the study. The weighted averages (WA) are notably larger as would be expected if ISV for the 2-digit industries (denoted below as 2-DI)

TABLE 2—DEPENDENT VARIABLE: EV

Equation[a]	Constant	MS	F/S	ISV	D	R^2
(2.1)	−4.965	−0.261	−0.482	0.304		0.203
[688]	(−17.20)	(−8.81)	(−6.07)	(6.67)		
(2.2)	−4.820	−0.250	−0.481	0.300	−0.088	0.204
[688]	(−14.78)	(−7.77)	(−6.06)	(6.55)	(−0.92)	
(2.3) $CR \leq 75$	−4.997	−0.263	−0.504	0.305		0.203
[667]	(−16.68)	(−8.51)	(−6.21)	(6.56)		
(2.4) $CR \leq 60$	−4.890	−0.256	−0.495	0.312		0.194
[593]	(−14.63)	(−7.78)	(−5.75)	(6.03)		
(2.5) $CR \leq 50$	−4.936	−0.276	−0.509	0.330		0.190
[491]	(−13.76)	(−7.12)	(−5.06)	(5.83)		
(2.6) $CR \leq 40$	−5.069	−0.292	−0.370	0.284		0.179
[405]	(−13.30)	(−6.88)	(−3.33)	(4.76)		
(2.7) $CR \leq 30$	−5.110	−0.299	−0.396	0.289		0.186
[296]	(−11.91)	(−6.00)	(−3.20)	(3.96)		

[a]See Table 1.

diversification differences, another explanatory variable was included. Variable (D) in equation (1.2) of Table 1 is the number (1,2,...,6) of 4-digit industries included in the list of each firm in Dun and Bradstreet's *Million Dollar Directory* (1981), except for $D = 6$ which means the number of industries is at least six. The estimated coefficient for D is negative, as expected, but it is clearly statistically insignificant. We do not take this to mean that diversification has no effect on sales variability. Instead we believe that the combined effect of including the variable ISV in the estimated equations and removing conspicuously diversified firms from our sample substantially reduces the observed diversification effect. We attribute the remaining strong inverse relationship between SV and MS to our hypothesis.

A second alternative explanation for our main result might be that large firms within industries—those with greater market shares —have market power supported by scale economies or other "mobility barriers;" and that these firms seek to take part of their profits in the form of avoiding uncertainty.[18] Since market shares often are taken to measure market power, this line of reasoning also would imply an inverse relationship between SV and MS. Equations (1.3)–(1.7) in Table 1 were estimated to ascertain the importance of market-power considerations. While firms' relative market share might be an index of market power in highly concentrated industries, it is less likely to be so in less concentrated industries. In industries with low degrees of concentration, the ability of firms to exert market power is limited so that differences in market share should not imply significant differences in market power.

To test whether the market share effect in equation (1.1) is a market-power effect, five subsamples of the original 856 firm sample were created. The first subsample contains firms in industries with 4-firm concentration ratios of 75 percent or less; the second subsample contains firms in industries with 4-firm concentration ratios of 60 percent or less; the remaining 3 subsamples contain firms in industries with 4-firm concentration ratios of at most 50, 40, and 30

reflects greater product diversification.

2-DI	ISV	WA ISV
2000	.0026	.0077
2200	.0078	.0123
2300	.0061	.0097
2400	.0092	.0114
2600	.0062	.0220
2800	.0071	.0108
3000	.0117	.0141
3600	.0121	.0153

[18]As mentioned earlier in fn. 11, this line of reasoning as well as the diversification effect have been offered for the inverse relationship that exists between rate-of-return variability and firm size.

percent, respectively. If the inverse relationship between SV and MS were capturing merely the effect of market power, then its strength should diminish for successively truncated subsamples. Clearly this does not happen; on the contrary the coefficients on MS remain negative and highly significant as the subsamples get smaller and less concentrated. Also, the R^2 statistics actually increase as industrial concentration falls. The robustness of the inverse relationship between SV and MS in even the 30 percent 4-firm concentration ratio subsample surely suggests that the market share effect is not due primarily to market power. We believe it supports our hypothesis that small, flexible rivals absorb a disaproportionate share of industrywide fluctuations.

The other testable implication of our theory mentioned above has to do with the utilization of labor. If small firms are more flexible than large ones, their use of labor, an input more variable than capital, should fluctuate more. Thus employment variability like sales variability should vary inversely with firm size.[19] To test this, a measure of firms' employment variability was computed. The natural logarithms of firm's annual employment levels were regressed onto a constant and a linear time trend, and adjustments for serial correlations were made. Then the standard error of these adjusted regressions was used as the measure of employment variability, EV. Since employment data for the period was not always available for firms in our sample of 856, the sample was reduced to 688 firms. Table 2 reports regressions analogous to all those found in Table 1 with EV replacing SV as the dependent variable. The results are quite similar although the R^2s are lower. The expected inverse relationship between EV and MS is quite strong, and making the previous adjustments for diversification and market-power considerations leaves it intact.

[19]This expectation is linked to another, that firm size and capital-labor ratios are positively associated. Using F as a measure of capital and E, average employment over 1970–80, as a measure of labor, MS and F/E have a correlation coefficient of .10 for our sample of 688 firms with a complete employment time-series.

III. Conclusion

The evidence presented here supports our theory of how output fluctuations are distributed among firms within industries. This theory holds that small firms are able to compete successfully with large, more static-efficient producers by absorbing a disproportionate share of industrywide output fluctuations. This is possible because small firms use production technologies that are more flexible than those chosen by large firms. Large rivals own the comparative advantage of lower minimum average costs, due largely to scale economies, while small competitors have an offsetting advantage in their superior responsiveness to cyclical or random swings in demand. It is noteworthy that this theory of industrywide structure with demand fluctuations is a generalized competitive equilibrium. The fact that fluctuations are not distributed equally or equiproportionately, and even the fact that small rivals absorb a disproportionate share, is perfectly compatible with competition. Our sensitivity tests fail to convince us that the observed pattern is better explained by market-power effects.

Insofar as this theory of industry structure is accurate and improves our understanding of observed intra-industry firm differences, it has implications for the interpretive usefulness of the structural variables that usually interest industrial organization economists. "Average" indicators of size, market share, etc. provide an incomplete picture of intra-industry differences in markets with significant demand fluctuations. In these markets especially it is important to consider specialized roles of technologically diverse firms. The shock-absorption role played collectively by smaller rivals is of special interest since it diminishes price fluctuations by increasing the elasticity of industries' supply response to fluctuating demand. A fluctuation-prone industry with only static-efficient firms would transmit greater price variability to buyers.

REFERENCES

Alexander, Sidney S., "The Effect of Size of Manufacturing Corporation on the Distri-

bution of the Rate of Return," *Review of Economics and Statistics*, August 1949, *31*, 229–35.

Caves, Richard E., "Uncertainty, Market Structure and Performance: Galbraith as Conventional Wisdom," in J. W. Markham and G. F. Papanek, eds., *Industrial Organization and Economic Development*, Cambridge: Harvard University Press, 1970, 283–301.

______ **and Pugel, T. A.**, *Intraindustry Differences in Conduct and Performance: Viable Strategies in U.S. Manufacturing Industries*, New York: New York University Press, 1980.

Davis, Hiram S., "Relation of Capital-Output Ratio to Firm Size in American Manufacturing: Some Additional Evidence," *Review of Economics and Statistics*, August 1956, *37*, 286–93.

Hall, Marshall and Weiss, Leonard, "Firm Size and Profitability," *Review of Economics and Statistics*, August 1967, *49*, 319–31.

Jacquemin, Alex and Saez, Wistano, "A Comparison of the Performance of the Largest European and Japanese Industrial Firms," *Oxford Economic Papers*, July 1976, *28*, 217–83.

Johnston, J., *Statistical Cost Analysis*, New York: McGraw-Hill, 1960.

Lucas, Robert E., "On the Size Distribution of Business Firms," *Bell Journal of Economics*, May 1978, *9*, 508–23.

McConnell, J. L., "Corporate Earnings by Size of Firms," *Survey of Current Business*, May 1945, *25*, 6–12.

Marcus, Matityahu, "Profitability and Size of Firm: Some Further Evidence," *Review of Economics and Statistics*, February 1969, *51*, 104–7.

Marschak, Thomas and Nelson, Richard R., "Flexibility, Uncertainty, and Economic Theory," *Metroeconomica*, April-August-December 1962, *14*, 42–60.

Mills, David E., "Demand Fluctuations and Endogenous Firm Flexibility," *Journal of Industrial Economics*, September 1984, *33*, 55–71.

Oi, Walter Y., "Heterogenous Firms and the Organization of Production," *Economic Inquiry*, April 1983, *21*, 147–71.

Porter, Michael E., "The Structure Within Industries and Companies' Performance," *Review of Economics and Statistics*, May 1979, *61*, 214–27.

Robinson, E. A. G., *The Structure of Competitive Industry*, Cambridge: Cambridge University Press, 1937.

Samuels, J. M. and Smyth, David J., "Profits, Variability of Profits and Firm Size," *Economica*, May 1968, *35*, 127–39.

Scherer, F. M., *Industrial Market Structure and Economic Performance*, Chicago: Rand McNally, 1980.

Sheshinski, Eytan and Dréze, Jacque H., "Demand Fluctuations, Capacity Utilization and Costs," *American Economic Review*, December 1976, *66*, 731–42.

Stekler, Herman O., "The Variability of Profitability with Size of Firm, 1947–1958," *American Statistical Association Journal*, December 1964, *59*, 1183–93.

Stigler, George J., "Production and Distribution in the Short Run," *Journal of Political Economy*, June 1939, *47*, 305–27.

______, *Capital and Rate of Return in Manufacturing Industries*, Princeton: Princeton University Press, 1963.

Whittington, Geoffrey, *The Prediction of Profitability*, Cambridge: Cambridge University Press, 1971.

Williamson, Oliver E., "Hierarchical Control and Optimum Firm Size," *Journal of Political Economy*, April 1967, *75*, 123–38.

Winn, Daryl N., "The Relations Between Rates of Return, Risk and Market Structure," *Quarterly Journal of Economics*, February 1977, *91*, 157–64.

Modeling the Formation of Price Expectations: A Bayesian Approach

By JOHN CASKEY*

Twice a year since 1947, Joseph Livingston, a financial columnist in Philadelphia, has recorded the inflation forecasts of professional economists. The surveyed economists generally underpredicted inflation from 1968 through 1979. This paper provides an empirical model of the formation of these economists' price expectations. The estimated model indicates that a convincing explanation for the Livingston inflation forecast series is that the economists had prior beliefs about the inflation process that they modified as they experienced the higher and more volatile inflation of the 1970's. This explanation is particularly appealing since it does not rely on statistically irrational behavior on the part of the forecasters.

A popular competing explanation for the underprediction of inflation during the 1970's argues that the Livingston forecasters either ignored available relevant information or processed that information poorly.[1] Many have noted that the Livingston survey participants had no monetary incentive to provide the best forecast possible.

While one cannot dismiss this alternative explanation, there are good grounds for seeking theories that do not rely upon either insufficient information sets or upon poor processing of available data. For example, commercial forecasters, with a monetary incentive to provide accurate forecasts and with large data banks, did not forecast significantly better than did the Livingston group. A comparison between the Livingston eight-month-ahead inflation forecasts and the comparable *CPI* inflation predictions of Data Resources, Incorporated (DRI) from 1973 to April, 1982 indicates the DRI's forecasts were no better than the mean forecast of the Livingston panel.[2] The mean absolute error of the DRI forecast was 3.22 percent while the mean absolute error of the Livingston forecasts was 3.2 percent. The DRI forecasts, the mean Livingston forecasts, and the actual

*Washington University, St. Louis, MO 63130. This paper was taken from research done for my dissertation at Stanford University. I am grateful for the advice that I received from the members of my dissertation committee: Robert E. Hall, Ben Bernanke, and Tom MaCurdy; as well as from a former fellow student, Tom Mroz. Many other people also improved this paper through their comments. I am particularly indebted to Don Cox, Paul Evans, Ed Greenberg, and two anonymous referees.

[1] For one example, see Douglas Pearce (1979).

[2] Jean Slamen, of the Federal Reserve Bank of Philadelphia, provided me with the Livingston data as well as with details on the methodology of the survey. In every survey, Livingston requests a forecast of the 6-month and 12-month-ahead *CPI* levels of a June or December starting date. On average, about 50 economists participate in each survey. They have diverse affiliations: the nonfinancial business sector, academic posts, financial institutions, and the Federal Reserve System. Livingston obtains a forecast by mailing the participants a one-page questionaire the month before the publication date. He publishes the survey results in June and December of each year. The questionaire includes the most current data available on the forecasts variables. In the case of the December publication of the *CPI* forecast, this is usually the September data but is sometimes the October data. John Carlson (1977) concludes that the participants base the forecasts on information that would be available one month before publication, i.e., in April or October. Before publication, Livingston sometimes adjusts the forecasts to take account of new information that has become available and that he thinks would have altered the forecasts of the participants. Rather than try to make similar adjustments or to use Livingston's published forecasts, I follow Carlson in working with the unadjusted forecasts and in treating them as 8 and 14-month forecasts.

My source of the DRI data is past issues of *The Data Resources Review of the U.S. Economy*, published from 1972 to the present. I am grateful to Katherine Kush of DRI for providing me with a number of these forecasts. Each issue of this monthly publication includes forecasts of the annual rates of change in the *CPI* for each quarter of the next year to year and a half. The DRI forecast that I used for comparison with the Livingston data was the forecast made at the end of April or October for the third and fourth quarters of the year, or first and second quarters of the next year, respectively. From these quarterly forecasts, I calculated the implicit 8-month DRI inflation forecast.

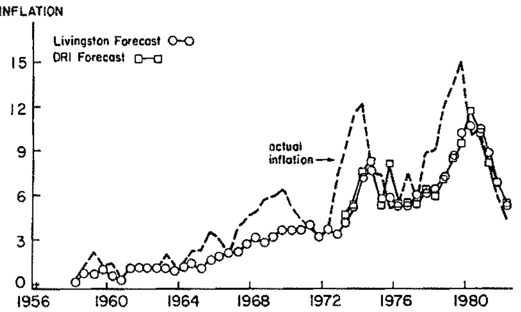


FIGURE 1. LIVINGSTON AND DRI FORECASTS

inflation rate are graphed in Figure 1. The similarity between the Livingston forecasts and those of DRI are remarkable. Since DRI is in the business of selling its forecasts, it has an obvious incentive to provide accurate forecasts. Furthermore, it had access to relevant data when making its forecasts. Moreover, these comparative observations about DRI's forecasts would apply equally well to the forecasts of the other major commercial forecasting establishments. Stephen McNees (1975; 1979; 1981) has done extensive comparisons of the performances of these forecasting services and has found that they all have similar records.

The model in this paper assumes that the Livingston forecasters processed new information in a statistically optimal manner given their beliefs about the inflation process. The statistical model is Bayesian since it represents the forecasters' subjective learning. Since I assume that the Livingston forecasts are the outcome of a specified forecasting and coefficient updating procedure, the missing element is the starting beliefs of the forecasters which would generate the observed series. I estimate the parameters of the Livingston economists' prior beliefs with maximum likelihood techniques.[3]

The results from this approach are very encouraging. I find that a credible set of 1958 prior beliefs about the inflation process combined with a Bayesian updating procedure accounts quite well for the Livingston forecasts. Moreover, the evolution of the coefficients over time agrees reasonably well with conventional ideas on the development of economists' beliefs about the determinants of inflation. For example, the estimated model indicates that the Livingston economists placed much more weight on recent inflation and monetary growth rates in predicting inflation at the close of the 1970's than they did in the early 1960's.

I. A Bayesian Learning Model of the Formation of Price Expectations

Before presenting a formal model of the formation of price expectations, it is helpful to first consider an informal description of the process. If asked to forecast the inflation rate in the United States over the next six months to a year, a group of economists might engage in the following thought process. From experience, reported research findings, and conventional wisdom, they believe that there is a set of variables, say, current inflation, monetary growth, fiscal policy, and aggregate demand that are good indicators of future inflation rates. They obtain information on the selected set of variables and on past correlations between these variables and subsequent inflation rates. The economists also ask themselves whether these observed past correlations should hold for

[3]This approach differs substantially from previous empirical studies of the formation of price expectations. I attempt to explain why the Livingston group forecast as it did. Previous research has focused on how the forecasts relate to other economic variables. Thus, while I explicitly model a learning and forecasting process, other researchers regress the Livingston price forecasts on variables that might influence price expectations. The estimated coefficients are then taken to identify the variables that the Livingston group used in forecasting prices and to indicate how much weight they placed on the variables. For example, see the papers by Stephen Turnovsky (1970), Jacob Frenkel (1975), K. Holden and D. A. Peel (1977), Kajal Lahari (1977), Rodney Jacobs and Robert Jones (1980), Donald Mullineaux (1980), David Resler (1980), Stephen Figlewski and Paul Wachtel (1981), and Edward Gramlich (1983). My comments on the methodology of this research must be qualified for three of these papers: Jacobs-Jones focus on the effect of changes in the price level, in the inflation rate, and in the acceleration of inflation; Mullineaux permits the coefficients to vary over time; and Figlewski-Wachtel obtain cross-sectional estimates over time.

the future. They might believe that the relationship between the selected variables and subsequent inflation rates has changed at a constant rate over time. If so, they give more weight to more recent correlations. They might think that sudden new economic policies, such as the imposition of sharp binding price controls, mean that old relationships contain no information relevant to future ones. In this case, they forecast on the basis of a best guess about future inflation without the benefit of observing correlations under this changed economic system. In all cases, the forecasters will not be bound by their initial beliefs. Rather, as they gain more experience with the economy, they may change their ideas about time-series relationships.

An outside observer trying to model the formation of expectations over time must then capture the major elements of above process. The Bayesian model in this paper takes an important step in that direction. It explicitly recognizes that the economists began forecasting with specific prior beliefs about the inflation process. It also permits the forecasters to learn over time and to modify their initial beliefs. In this implementation of the Bayesian learning mode, however, I impose an important restriction. I assume that the forecasters believed the relationship between the variables in the information set and subsequent inflation rates to be stable over the relevant period. While the validity of this assumption may be challenged, it is nevertheless worthwhile to begin with this restricted model and to see how well it can explain the observed price expectations on its own.[4]

The Bayesian model of the formation of Livingston price expectations has four elements: the inflation process in which the group believed, the group's prior beliefs about the parameters of that process, the statistical rule that the group used in updating its beliefs about the coefficients, and the group's forecasting rule. In this section, these elements are specified. In the next section, from the observed Livingston forecasts and based on an assumed information set, I estimate the parameters of the Livingston group's prior beliefs about the coefficients in the inflation process.

The model of the formation of price expectations assumes that the Livingston group believed that the one-period-ahead inflation rate had the time-series representation

$$(1) \qquad y_t = X_{t-1}B + e_t \qquad e \sim N(0,\sigma^2)$$

where y is the actual inflation rate, X is the k element row vector of observations on the right-hand side variables which includes lagged dependent variables, and e is the disturbance term.[5] The y_0 is assumed to be fixed. The time subscript, $t-1$, indicates that these are variables that could have been known in the earlier period. Thus X_{t-1} could also include variables from two periods before or more.

A multivariate normal distribution for the coefficients in equation (1) represents the initial beliefs of the Livingston forecasters about the inflation process. It is convenient to write the prior distribution for B as

$$(2) \qquad B \sim N(B_0, \sigma^2 V_0^{-1}),$$

and $\sigma^2 V_0^{-1} \equiv Q^{-1}$. In what follows, V_0^{-1} is referred to as the prior covariance matrix, although it should be remembered that it is actually merely proportional to the true co-

[4]A completely general model with unrestricted parameter variation cannot be estimated. One could, however, somewhat relax the stability assumption. In my dissertation, I modify the model presented here in order to permit one abrupt change in the Livingston group's prior as a consequence of new information from outside of the information set. I used this modified model in a study of the impact on the group's price expectations of the Fed's October 1979 announcement. However, with semiannual price forecast observations and with an event so near the end of the sample period, it was necessary to impose severe and probably distorting restrictions on the parameters in order to obtain enough degrees of freedom to estimate them. Alternatively, one could permit a smooth discounting of the information content of old data. One cannot, however, estimate both this rate of discount as well as the prior covariance matrix for the parameter B without imposing arbitrary restrictions.

[5]The presence of lagged dependent variables in the information set justifies serial independence. One could, however, estimate the model with a serially correlated error term, though this would add several parameters.

variance matrix, Q^{-1}. The B_0 is the Livingston group's prior beliefs about the value of B. The matrix V_0^{-1} is a measure of their uncertainty in their prior beliefs relative to σ^2.

I assume that the Livingston group forecast inflation on the basis of their evolving beliefs about B. The Kalman filter, which in its usual formulations assumes σ is known, is the recursive statistical rule that represents the updating of beliefs about B. As Richard Meinhold and Nozer Singpurwalla (1983) and others have explained, if σ is known, the Kalman filter is a convenient method for obtaining the parameters of the normally distributed posterior distribution on B conditional on new y and X observations.

As is shown generally in Arnold Zellner (1971) (or, for this specific case, in my dissertation), although σ is not known, one can employ the Kalman filter to represent the Livingston group's updating of their beliefs about B. If σ is unknown, one models a Bayesian learning rule by specifying a prior distribution for σ and deriving a joint posterior distribution for B and σ conditional on the new y and X observations. Assuming that σ has an inverted *gamma* distribution, the marginal posterior distribution for B will be multivariate t-statistic. The mean, B_m, of this posterior distribution represents the Livingston group's new beliefs about B. This mean is equal to the mean of the normal distribution on B conditional on a known σ.

Each period the Livingston group obtains new information on inflation and on other variables in its information set. They use this new information to update their beliefs about B according to the three recursive equations of the Kalman filter:

$$(3) \quad \hat{B}_t = \hat{B}_{t-1} + K_t\left(y_t - X_{t-1}\hat{B}_{t-1}\right)$$

$$(4) \quad \hat{V}_t^{-1} = \hat{V}_{t-1}^{-1} - K_t X_{t-1}\hat{V}_{t-1}^{-1}$$

$$(5) \quad K_t = \sigma^2\hat{V}_{t-1}^{-1}X'_{t-1} \times\left[X_{t-1}\sigma^2\hat{V}_{t-1}^{-1}X'_{t-1} + \sigma^2\right]^{-1} = \hat{V}_{t-1}^{-1}X'_{t-1}\left[X_{t-1}\hat{V}_{t-1}^{-1}X'_{t-1} + 1\right]^{-1},$$

where $\hat{B}_{t-1}$ and $\hat{V}_{t-1}$ are the Livingston group's beliefs in period $t-1$ about B and a measure of its uncertainty relative to σ^2. The $\hat{B}$ is a k element column vector. The $\hat{V}$ is a $k \times k$ matrix and K is the k element column vector known as the Kalman gain.

Assuming that the Livingston group has a quadratic loss function, its inflation forecast is

$$(6) \quad yp_{t+1} = X_t\hat{B}_t.$$

One observes the Livingston predictions, yp. If it is assumed that the predictions are the outcome of the Bayesian coefficient updating procedure specified in equations (3)–(5), the only missing elements behind the formation of the group's expectations are the group's information set, X, and the group's starting beliefs about B. If one makes reasonable assumptions about the components of X, one can then estimate the starting beliefs of the Livingston group about the inflation process.[6]

In empirically implementing this model of the formation of the Livingston group's price expectations, the individual's thought processes cannot, of course, be fully replicated. First, I will likely omit some of the minor variables used in their forecasting process. Second, I estimate the model using the revised data that is now available, not the data that was available to the Livingston panel at the time of each forecast. Third, the model assumes that the Livingston group believed B in equation (1) to be constant. In sum, these factors introduce an error term into the equality between the observed Livingston forecasts and the forecasts that my replication of their expectation process produces.

[6]As discussed in my dissertation, one cannot estimate the parameters of a prior distribution on σ since one does not observe the variance of the distribution on the y forecasts. One could, however, hypothesize that the variance across the survey respondents in each period bears a particular relation to the variance on the posterior distribution of y, where y_t is distributed $N(X_{t-1}\hat{B}_{t-1}, X'_{t-1}V_t^{-1}X_{t-1} + w^2)$. Unfortunately, it is unclear what the particular relation would be. Moreover, such an approach is philosophically at odds with my implicit treatment of the Livingston forecasters as a homogeneous group.

Therefore, I rewrite (6) as

$$(7)\qquad yl_t = yp_t + u_t \qquad\qquad u \sim N(0, \sigma_u^2),$$

where yl is the observed Livingston inflation forecast. The error term is assumed to be normally distributed and serially uncorrelated: we expect that the impact of omitted elements would be roughly constant over time, and the imperfect modeling introduces no clear bias into the error.

The estimation procedure seeks the B_0 and V_0^{-1} that maximize the likelihood of the log-likelihood function based on (7), which is

$$(8)\quad \operatorname{Log} L = (-n/2)\log(\sigma_u^2) - (1/2\sigma_u^2)(yl - yp)'(yl - yp) + \text{constant}.$$

With four variables in X, however, this involves estimating fifteen parameters. Consequently, to gain a few additional degrees of freedom without unduly restricting the model, I assume that the precision matrix, V_0, can be parameterized by a one-factor scheme. That is, I rewrite the symmetric precision matrix V_0 in the following manner:

$$\begin{bmatrix} V_{11} & V_{12} & V_{13} & V_{14} \\ & V_{22} & V_{23} & V_{24} \\ & & V_{33} & V_{34} \\ & & & V_{44} \end{bmatrix} = f'f + D$$

where $\quad f = (f_1, f_2, f_3, f_4)$

and $\quad D = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & d_3 & \\ & & & d_4 \end{bmatrix}$

The restriction permits a full covariance matrix, yet reduces the number of parameters in the model by two.[7] With four right-hand side variables composing X, there are thirteen parameters to estimate: the four prior means, the four elements of the diagonal matrix, the four factor loadings, and the variance σ_u^2.

II. Estimates of the Parameters of the Model

In estimating the model, I assume that the Livingston panel bases its prediction of the eight-month-ahead annual inflation rate on the following information set: a constant, the past ten-month's annual inflation rate (*INFL*), the annual rate of change in the past six-month money stock (*DM1L*), and the past quarter's federal budget surplus as a percentage of *GNP* (*SURL*).[8] I also experimented with several alternative specifications of the information set. I replaced the fourth variable, *SURL*, with the full-employment federal budget surplus as a percentage of potential *GNP* (*FSURL*) and with the unemployment rate (*UNL*). I also estimated the model with five variables in the information set: the original four along with the unemployment rate.

I might summarize the estimation technique by stating that the maximum likelihood procedure chooses values for the parameters of the Livingston group's prior distribution on B such that, when this prior is updated each period by the Kalman filter and used to forecast, the forecast series resembles the Livingston series.

I estimate the model using data from 1959 through 1982. Thus, the estimated parameters are for the prior distribution that the Livingston group placed on B in the second half of 1958. This prior distribution represents the starting beliefs of the Livingston panel as to the time-series relationship between the four selected right-hand side variables and the future inflation rate. Since the model specifies the manner in which the forecasters would update their beliefs, the estimates tell us both their starting beliefs and the evolution of those beliefs over time. As the forecasters obtain additional observations on the inflation process, they "learn," or, equivalently, update their beliefs about B.

[7] This restriction is particularly important when I estimate the model with 5 right-hand side variables. In this case it reduces the number of parameters from 21 to 16. With 48 observations this reduction can be important. Moreover, the restriction is minimal in that it still permits each element of the covariance matrix to be different.

[8] See the Appendix for details on the variables.

TABLE 1—ESTIMATES OF THE BAYESIAN LEARNING MODEL:1958:2

	Constant	*INFL*	*DM1L*	*SURL*	*FSURL*	*UNL*
Prior Means[a]	0.283	0.093	−0.17	−0.572		
	(.042)	(.0086)	(.019)	(.067)		
Prior Covariance Matrix[b]						
Constant	.312	.00005	.0413	−.0132		
INFL		.00277	−.00006	.000018		
DM1L			.00702	.0144		
FSURL				.169		
Error Variance Estimate − σ_u^2	.075	(.0022)				
Likelihood Function Value −38.15						
Alternative Specifications of the Information Set						
Prior Means	.876	.135	−.092		−.37	
	(.083)	(.0093)	(.0275)		(.073)	
Likelihood Function Value −26.73						
Prior Means	.33	.042	−.10			−.028
	(.185)	(.0127)	(.0148)			(.035)
Likelihood Function Value −27.96						
Prior Means	.92	.052	−.176	−.49		−.105
	(.16)	(.016)	(.028)	(1.89)		(.17)
Likelihood Function Value −39.40						

[a]Standard errors of the estimates are shown in parentheses.
[b]This covariance matrix is constructed from the estimated factor loadings and the elements of the diagonal matrix: f_1 −54.1 (42.5); f_2 20.8 (19.6); f_3 384.4 (300.6); f_4 −40.3 (31.9); d_1 2.34 (0.47); d_2 350.1 (11.1); d_3 42.3 (26.6); d_4 4.9 (1.08)

Employing a program for estimating the parameters of nonlinear equations, I obtained the maximum likelihood estimates presented in Table 1. The estimated prior means are those that the Livingston group placed on the coefficients of the reduced-form inflation model (1) in late 1958. The prior covariance matrix is estimated up to an unknown scale term σ^2, and indicates the relative confidence that the forecasters had in their beliefs about the respective coefficients and in their beliefs about the interaction among these coefficients. Finally, the standard errors indicate the econometrician's uncertainty about the estimates and are unrelated to the Livingston panel's beliefs.

How should one interpret these estimates? The estimated prior mean for the coefficient on *INFL* indicates that, had the inflation rate previous to the 1959:1 Livingston forecast been 1.14 percent instead of the actual 0.14 percent, the group would have increased its inflation forecast from 0.61 to 0.62 percent. In other words, prior to seeing the 1959 data, the Livingston panel thought that changes in the previous period's inflation rate contained little information relevant to the next period's inflation rate.

The estimated prior mean for the coefficient on the rate of change in the money stock indicates that, had *M*1 grown at 1.72 percent a year in late 1958 instead of the actual 0.72 percent growth rate, the Livingston panel would have reduced its 1959:1 inflation forecast from 0.61 to 0.44 percent. Moreover, the negative sign on this coefficient remained when *SURL* was replaced with unemployment, *UNL*, or the full-employment deficit measure, *FSURL*. However, in both of these cases the estimated prior mean increased from −.17 to about −.10. Although most economists today would place a positive coefficient on *DM1L*, there is evidence that these estimates of the beliefs of economists in 1959 may be reasonable. Benjamin Klein (1975, 1978), for example, has argued that the United States was at that time on an unofficial gold standard. Thus, money supply increases were likely to be reversed in later periods and did not signal a relaxed monetary policy. Nevertheless, while a zero coefficient would not be

surprising for forecasters of 1959, a negative one is surprising.

Finally, the prior coefficient on *SURL* indicates that, had this measure of the budget surplus as a percentage of *GNP* been 1.6 percent instead of the actual .6 percent in the third quarter of 1958, the Livingston group would have changed their inflation forecast from .61 to .04 percent. Thus, the panel of 1959 believed that changes in this variable contained substantial information about the next period's inflation rate. This finding supports the general notion that economists in 1959 placed much weight on fiscal policy as a determinant of aggregate demand and, consequently, of inflation. Further support for this interpretation lies in the estimated prior mean of $-.37$ if *SURL* is replaced by *FSURL*, the full-employment surplus as a percentage of potential *GNP*. Similarly, when the model was reestimated with a fifth right-hand side variable, unemployment (*UNL*), the estimated prior mean on *SURL* was $-.48$, indicating that the finding is fairly robust, even after controlling for the state of aggregate demand.

The estimates of the 1958:2 prior covariance matrix shown in Table 1 reveal the relative confidence that the Livingston group placed in their beliefs about the value of the coefficients. Since σ^2 is unknown, the absolute size of the estimated coefficients is indeterminate.[9] The relatively loose prior variance on the constant suggests the panel's willingness to quickly revise their beliefs about the value of this coefficient. To a lesser extent, the same holds true for the coefficient on the federal budget surplus. The prior variances on the other two variables are of about the same magnitude and are much tighter than that on the constant or on the *SURL* coefficient. Thus, the panel would be much slower to revise its beliefs about these coefficients in the face of data that indicates a different time-series relationship between these variables and future inflation.

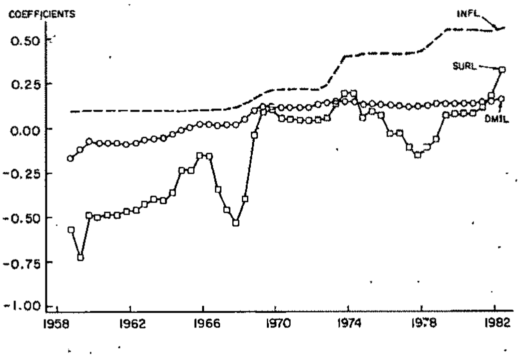


FIGURE 2. EVOLUTION OF THE COEFFICIENTS

The strength of the Bayesian learning approach lies not only in its ability to estimate the starting beliefs of the Livingston panel, but also in its ability to capture the evolution of those beliefs over time. Figure 2 shows the evolving beliefs of the Livingston panel about the time-series relationship among the variables in their information set and future inflation. Examination of the graph shows that the panel learned to place more and more weight on the previous period's inflation rate in forecasting future prices. A second striking pattern is the change in importance given to money growth: panel members steadily increased the weight that they placed on past changes in money in forecasting prices.[10]

Particularly interesting is the evolution of the coefficient on *SURL*. The smooth decline in the weight that the group placed on this variable agrees with the popular notion that as economists began to place more weight on monetary growth in forecasting inflation, they

[9]One possible estimate of a ball park figure for the panel's 1958:2 σ^2 is obtained by regressing y_t on X_{t-1} from 1953:1 through 1958:1 and using the residuals to construct a $\hat{\sigma}^2$. This approach yields a $\hat{\sigma}^2$ of approximately 0.4.

[10]Mullineaux obtained a similar result using a varying parameter regression technique over the period from 1966 to 1977. There is, however, a major difference between our two approaches. Mullineaux regresses the Livingston forecasts on variables that one might reasonably assume to enter the Livingston group's information set. He permits the coefficient to vary over time in a restricted manner. While this technique does show that the relation between the Livingston forecasts and the assumed set of information variables shifted over time, it does not provide any grounds, such as a learning process, for why the shift occurred. Consequently, one should not construe Mullineaux's noteworthy study of expectations to be a model of the formation of expectations in the sense of explaining why people forecast as they did.

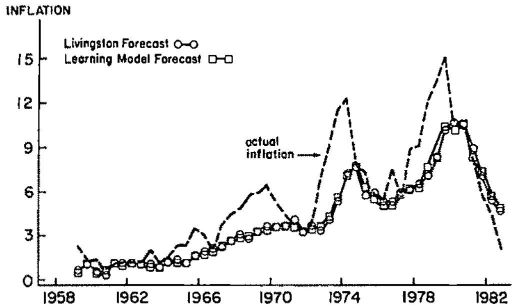


FIGURE 3. LIVINGSTON AND LEARNING MODEL FORECASTS

placed less importance on fiscal policy. The post-1969 positive coefficient on *SURL* is, however, a puzzle. One thought is that since part of fiscal policy is an automatic response to business cycle conditions, *SURL* may proxy both discretionary fiscal policy as well as aggregate demand conditions. When, however, *SURL* is replaced with *FSURL*, the forecasting coefficient still becomes positive post-1977.

As Figure 3 shows, the Bayesian learning model explains quite well the Livingston group's forecasts. The forecasts from the estimated prior and the updated coefficients coincide with the Livingston group's forecasts to an impressive degree. The mean squared difference between the two forecast series is 0.075. This settles one issue conclusively: a specific set of 1958:2 prior beliefs for the Livingston panel exists that, when combined with a Bayesian model of learning, accounts quite well for the Livingston forecasts.

III. Conclusion

This paper demonstrates that one can explain the Livingston inflation forecasts as the product of a learning process. Two underlying assumptions should, however, be emphasized. It is assumed, not tested, that the panel followed Bayes' Rule in updating their beliefs. It is also assumed that the panel believed the underlying parameters of the inflation process were constant over the estimation period. Further research could perhaps adopt a more general learning model that permits either a continuous or discontinuous discounting of the information content of older data.

We should not interpret the findings presented in this paper as proof that the Livingston forecasts were the most accurate forecasts possible given available information. The estimates do show, however, that the Livingston forecasts are consistent with optimal forecasting behavior. This is because I find that plausible initial beliefs combined with an updating procedure that follows the laws of conditional probability account for the observed forecasts. We do not need to appeal to statistical irrationality in explaining the Livingston price expectations.

APPENDIX

INFLV: The mean of the inflation forecasts (*CPI*) of the participants in the Livingston survey.

INF: The *CPI*-based annual inflation rate over the eight-month period that coincides with the horizon of the Livingston forecast (i.e., April/December or October/June). The *CPI* monthly series was obtained from the U.S. Department of Commerce *Business Conditions Digest*. The index is the All Items index. The inflation rate is calculated according to the following formula:

$$INF=100\left(\exp\left(\frac{12}{8}\log\frac{CPI_t}{CPI_{t-8}}\right)-1\right).$$

INFL: A similarly calculated *CPI* annual inflation rate over the ten-month period just prior to the date of the Livingston forecasts (i.e., June/April or December/October).

DM1L: The annual growth rate in the money stock (*M*1) over the six-month period just prior to the date of the Livingston forecast (i.e., October/April or April/October). The monthly data on *M*1 was provided by the Federal Reserve Board. The annual rate of growth was calculated according to the following formula:

$$DM1=100\left(\exp\left(\frac{12}{6}\log\frac{M1_t}{M1_{t-6}}\right)-1\right).$$

SURL: The federal government's seasonally adjusted quarterly budget surplus at an

annual rate divided by the seasonally adjusted quarterly measure of *GNP* at an annual rate. This ratio was then multiplied by 100 in order to scale it similarly to the other variables. The data is for the quarter just prior to the Livingston forecast date (i.e., the first and third quarters). The source of both measures is the U.S. Department of Commerce, *National Income and Product Account of the United States.*

FSURL: This variable is constructed similarly to *SURL* but using the full-employment federal surplus and potential *GNP*. The Federal Reserve Bank of St. Louis provided me with this data.

UNL: Unemployment as a percentage of all workers in the months of April or October. The data source is the Bureau of Labor Statistics.

REFERENCES

Carlson, John A., "A Study of Price Forecasts," *Annals of Economic and Social Measurement*, 1977, *6*, 27–56.

______, "Perceptions (or Misperceptions) of Inflation," in William Gale, ed., *Inflation: Causes, Consequences, and Control*, Cambridge: Oelgeschlager, Gunn, and Hain, 1981.

Caskey, John, "Modeling the Formation of Price Expectations," unpublished doctoral dissertation, Stanford University, 1983.

Figlewski, Stephen and Wachtel, Paul, "The Formation of Inflationary Expectations," *Review of Economics and Statistics*, February 1981, *63*, 1–10.

Frenkel, Jacob A., "Inflation and the Formation of Expectations," *Journal of Monetary Economics*, June 1975, *1*, 403–21.

Gramlich, Edward M., "Models of Inflation Expectations Formation: A Comparison of Household and Economist Forecasts," *Journal of Money, Credit and Banking*, May 1983, *15*, 155–73.

Holden, K. and Peel, D. A., "An Empirical Investigation of Inflationary Expectations," *Oxford Bulletin of Economics and Statistics*, November 1977, *39*, 291–99.

Jacobs, Rodney L. and Jones Robert A., "Price Expectations in the United States: 1947–75," *American Economic Review*, June 1980, *70*, 269–77.

Klein, Benjamin, "Our New Monetary Standard: The Measurement and Effects of Price Uncertainty, 1880–1973," *Economic Inquiry*, December 1975, *13*, 461–84.

______, "The Measurement of Long- and Short-term Price Uncertainty: A Moving Regression Time Series Analysis," *Economic Inquiry*, July 1978, *16*, 438–52.

Lahari, Kajal, "A Joint Study of Expectations Formation and the Shifting Phillips Curve," *Journal of Monetary Economics*, July 1977, *3*, 347–57.

McNees, Stephen K., "An Evaluation of Economic Forecasts," *New England Economic Review*, November/December 1975, 3–39.

______, "The Forecasting Record for the 1970s," *New England Economic Review*, September/October 1979, 33–53.

______, "The Recent Record of Thirteen Forecasters," *New England Economic Review*, September/October 1981, 5–21.

Meinhold, Richard and Singpurwalla, Nozer, "Understanding the Kalman Filter," *American Statistician*, May 1983, *37*, 123–27.

Mullineaux, Donald J., "Inflation Expectations and Money Growth in the United States," *American Economic Review*, March 1980, *70*, 149–61.

Pearce, Douglas K., "Comparing Survey and Rational Measures of Expected Inflation," *Journal of Money, Credit and Banking*, November 1979, *11*, 447–56.

Resler, David H., "The Formation of Inflation Expectations," *Federal Reserve Bank of St. Louis Review*, April 1980, *62*, 2–12.

Turnovsky, Stephen J., "Empirical Evidence on the Formation of Price Expectations" *Journal of the American Statistical Association*, December 1970, *65*, 1441–54.

Zellner, Arnold, *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley & Sons, 1971.

Data Resources, Incorporated, *The Data Resources Review of the U.S. Economy*, Lexington: DRI, various months.

U.S. Department of Commerce, *Business Conditions Digest*, Washington: USGPO, various months.

______, *National Income and Product Account of the United States, 1929–1776*, Washington: USGPO, 1981.

______, *Survey of Current Business*, Washington: USGPO, various months.

Precautionary Saving and Accidental Bequests

By ANDREW B. ABEL*

This paper presents a simple general equilibrium model of precautionary saving and accidental bequests. This model is used to analyze the implications of individual lifetime uncertainty for aggregate consumption and capital accumulation. A precautionary demand for saving arises because an individual consumer does not know in advance the date at which he will die, and he wants to avoid low levels of consumption in the event that he lives longer than expected. An implication of this precautionary saving is that when death does occur, the consumer is generally holding some wealth, which is then passed on to his heirs in the form of an accidental bequest. Even if all consumers have the same *ex ante* mortality probabilities, there will be some intracohort variation in the date of death; consequently there will be a nondegenerate distribution of bequests left by consumers in a cohort. This nondegenerate distribution of bequests left by one generation induces variation in the distributions of wealth, consumption, and bequests of subsequent generations.

The importance of bequests in aggregate saving has been established by Laurence Kotlikoff and Lawrence Summers (1981) who reported that 80 percent of U.S. household wealth is inherited wealth. One interpretation of this finding is that the simple life cycle model without bequest motives is an inadequate description of saving behavior in the United States, but the model I present demonstrates that accidental bequests by selfish consumers can account for a potentially sizeable fraction of aggregate wealth. Although some part of bequests, especially by the wealthy, undoubtedly results from an explicit bequest motive, accidental bequests also play a role in the intergenerational transfer of wealth as well as in the intragenerational variation in wealth. In order to focus on the role of accidental bequests, I purposely exclude a bequest motive from the specification of the utility function.[1]

The effects of lifetime uncertainty on individual consumption behavior were first examined formally in a seminal paper by Menachem Yaari (1965). Yaari's model provided the basic framework for virtually all subsequent work on uncertain lifetimes including well-known papers by Nils Hakansson (1969), Stanley Fischer (1973), Robert Barro and James Friedman (1977), David Levhari and Leonard Mirman (1977), and Kotlikoff and Avia Spivak (1981). However, all of these papers focused on the consumption decision of an individual and ignored the effect of accidental bequests on the behavior of the recipients of these accidental bequests.[2] As will be shown at various points in this paper, changes in the economic environment can have effects on aggregate behavior which differ sharply from the effects on individual behavior because of the endogenous adjustment of bequests. The effects of

*Department of Economics, Harvard University, Cambridge, MA 02138. This paper is a revised version of "Bequests and Social Security with Uncertain Lifetimes," NBER Working Paper No. 1372, June 1984. I thank Olivier Blanchard, John Burbidge, Christopher Cavanagh, Robert Clower, Stanley Fischer, Benjamin Friedman, Elhanan Helpman, Mervyn King, Nobuhiro Kiyotaki, Laurence Kotlikoff, Robert McDonald, James Pesando, Julio Rotemberg, Lawrence Summers, Lars Svensson, Mark Watson, and Jeffrey Wolcowitz for valuable conversations and correspondence. I also thank the participants in seminars at Boston University, Columbia University, Cornell University, Federal Reserve Board, Harvard University, University of Iowa, MIT, University of Minnesota, University of Montreal, National Bureau of Economic Research, North Carolina State University, the Wharton School, and Yale University, and two anonymous referees for their helpful comments. I also thank the National Science Foundation for financial support.

[1] In the presence of perfect annuity and life insurance markets, there may or may not be bequests (depending on the presence or absence of a bequest motive), but there would be no accidental bequests.

[2] Kotlikoff and Spivak focus on the role of the family in providing an (incomplete) annuities market, but stop short of a full-scale overlapping generations model in which the distribution of bequests is determined endogenously.

the endogenous adjustment of bequests are modeled by embedding consumers with uncertain lifetimes into an overlapping-generations model à la Franco Modigliani and Richard Brumberg (1979), Paul Samuelson (1958), and Peter Diamond (1965). The model is dramatically different from the overlapping generations model with uncertain lifetimes proposed by Eytan Sheshinski and Yoram Weiss (1981), because Sheshinski-Weiss assume that all consumers who are born at the same date also die on the same date. Thus, unlike the model presented below, the Sheshinski-Weiss model does not generate intracohort variation in bequests, consumption, and wealth.

Zvi Eckstein, Martin Eichenbaum, and Dan Peled (1985a) have developed an overlapping generations model in which consumers have identical *ex ante* mortality probabilities but die at different ages. Since the Eckstein et al. model, which was developed independently of my model presented below, is similar to that model, it is worth commenting on the differences between the two models. First, and most importantly, the Eckstein et al. model has no capital although one could interpret that model as applying to an economy in which the net rate of return on capital is zero (i.e., a costless storage technology). However, as shown below, the effects of Social Security policy differ depending on whether or not the rate of return on capital is zero. Second, in my model, the instantaneous utility function is assumed to display hyperbolic absolute risk aversion (*HARA*), whereas Eckstein et al. use a more general concave utility function. However, their formulation is not as general as it might first appear because Eckstein et al. must at some point assume that the concavity of the derived saving function is "not too large" without presenting the implied restrictions on the utility function. An advantage of the *HARA* utility function used here is that it leads to linear decision rules, thereby making the analysis easily tractable. Third, my model presented below allows for nonzero rates of time preference and population growth, whereas each of these rates is assumed to be zero by Eckstein et al.

In Section I, I present a simple model of individual consumption behavior in the presence of an uncertain lifetime. In Section II, I trace the effects of accidental bequests on the saving and consumption of subsequent generations and calculate the steady-state intracohort distributions of consumption, wealth, and bequests. The next three sections analyze the aggregate and distributional consequences of introducing different types of annuities into the economy. Section III demonstrates that, in the absence of an annuity market, the introduction of a fully funded actuarially fair Social Security system leads to a reduction in the steady-state national capital stock. In addition, the introduction of actuarially fair Social Security reduces all central moments of the distributions of consumption, wealth, and bequests. Section IV is devoted to an analysis of the transition path to the new steady state. In Section V, I show that the introduction of a competitive annuity market can cause the steady-state capital stock either to rise or fall depending on whether the coefficient of relative risk aversion is less than or greater than a certain critical value. Concluding remarks and directions for further research are presented in Section VI.

I. Individual Consumption Behavior under Uncertain Lifetime

Consider an economy with many consumers and a single commodity. This commodity can be either consumed or invested. If one unit of the commodity is invested, it yields R units of the commodity in the following period. Each consumer lives either one or two periods. A consumer works during the first period of his (or her) life earning a fixed labor income Y.[3] Also in the first period of his life, a consumer consumes an amount c_1 and pays a tax T. At the end of the first period of his life, the consumer has $G \geq 1$ children. There is a probability p that the consumer dies at the end of his first

[3] It is assumed that the production function is linear in capital and employment. Let N_t be the number of consumers born at the beginning of period t and let K_{t-1} be the average capital stock held at the end of period $t-1$ by consumers born at the beginning of period $t-1$. Then aggregate output in period t is $N_tY+RN_{t-1}K_{t-1}$.

period of life[4] (after having the children). If the consumer survives to the second period of life, he does not work but receives a Social Security payment S. He then consumes an amount c_2. When a consumer dies (either at the end of period one or period two), any unconsumed wealth is divided equally among his children.

Each consumer chooses c_1 and c_2 to maximize the following utility function

$$U(c_1)+(1-p)\delta U(c_2), \tag{1}$$

where $0<\delta\leq 1$. This utility function is based on the uncertain lifetime literature in which the discounted utility index for period j is multiplied by the probability of being alive in period j. This formulation is simply the expected value of a state-contingent utility function in which $U(c_j)$ is the utility index contingent on being alive at age j, and the utility index is identically zero contingent on not being alive at age j.[5] According to the utility function in (1), consumers do not care about their children; they derive no utility from leaving bequests.[6]

Up to this point it may appear that all consumers are identical: they have identical utility functions, labor income Y, taxes T, childbearing characteristics, probabilities of survival, and, if they survive, identical Social Security benefits S. However, different consumers receive bequests of different sizes depending on the mortality history of the earlier generations of their families. Let B be the bequest a consumer receives from his parent when he is born.[7] For the moment, take B as given; the determination of B will be discussed in Section II.

Finally, W is defined to be the wealth held by a consumer at the end of the first period of his life:

$$W=B+Y-T-c_1. \tag{2}$$

If a consumer dies at the end of his first period of life, each of his children receives RW/G as a bequest at the beginning of the following period. If the consumer survives into the second period, he consumes $c_2=RW+S$, because he derives no utility from leaving a bequest. Using equation (2), we have

$$c_2=R[B+Y-T-c_1]+S. \tag{3}$$

The consumer's first-period consumption is determined by maximizing (1) with respect to c_1, subject to the constraint in (3), to obtain

$$U'(c_1)=(1-p)R\delta U'(c_2). \tag{4}$$

The first-order condition in (4) equates the marginal utility of a unit of first-period consumption with the expected present value of the utility from the R units of second-period consumption which could be obtained by reducing current consumption by one unit. I specify the utility index $U(c)$ to be a member of the *HARA* (hyperbolic absolute risk aversion) family[8]

$$U(c)=\frac{1-\gamma}{\gamma}\left(\frac{\beta c}{1-\gamma}+\eta\right)^{\gamma}. \tag{5}$$

With *HARA* utility, the optimal value of

[4] Although individual consumers face uncertainty about their date of death, there is no aggregate uncertainty; a fraction p of consumers in each generation dies at the end of the first period of life.

[5] It is not necessary that the utility index is equal to zero in the case of death. All that is required is that utility in the state of death does not depend on the level of wealth.

[6] Fischer and Sheshinski-Weiss model consumers as deriving utility from leaving a bequest. This utility is a function of the size of the bequest. Barro (1974) and Allan Drazen (1978), in models without lifetime uncertainty, assume that consumers derive utility from the utility of their children. Douglas Bernheim, Andrei Shleifer, and Summers (1984) argue that parents use the prospect of bequests as a way to induce their children to behave in ways that the parent wants. Thus, although parents care only about their own utility, they find it optimal to leave bequests.

[7] If a parent dies after the first period of his life, his child receives a bequest B at the beginning of the first period of the child's life. If a parent lives two periods, then as shown below, the child receives no bequest in either period; in this case, of course, the bequest received at birth by the child is zero.

[8] The utility function in (5) is subject to the following restrictions: $\gamma\neq 1$; $\beta>0$; $(\beta c/(1-\gamma))+\eta>0$; $\eta=1$ if $\gamma=-\infty$. The *HARA* family of utility functions includes the following special cases: (i) constant relative risk aversion ($\eta=0$), which includes logarithmic utility if $\gamma=0$; (ii) constant absolute risk aversion ($\gamma=+\infty$); and (iii) quadratic utility ($\gamma=2$).

first-period consumption can be written as a linear function of the present value of lifetime income $B+Y-T+R^{-1}S$,

$$(6a)\quad c_1 = a(B+Y-T+R^{-1}S)+b,$$

where

$$(6b)\quad 0<a = \left[1+R^{-1}[(1-p)R\delta]^{1/(1-\gamma)}\right]^{-1}<1$$

$$(6c)\quad b=((1-\gamma)\eta/\beta)aR^{-1} \times\left[1-((1-p)R\delta)^{1/(1-\gamma)}\right].$$

Note that a, the marginal propensity to consume, is a positive constant less than one.

If $U(c)$ has constant relative risk aversion ($\eta=0$), then $b=0$ and first-period consumption is proportional to the present value of disposable lifetime resources $B+Y-T+R^{-1}S$. Let $\sigma\equiv 1-\gamma$ be the (constant) coefficient of relative risk aversion and note that if $R=\delta=1$ (i.e., zero time preference and zero net rate of return on capital), then the fraction of total disposable resources $(B+Y-T+S)$ consumed in the first period of life is $a=[1+(1-p)^{1/\sigma}]^{-1}$. The greater the coefficient of relative risk aversion, the smaller the fraction of disposable resources consumed in the first period. In the limit as $\sigma\to\infty$, a consumer would consume one-half of disposable resources in the first period; with a large σ, the desire to smooth consumption is so strong that the consumer is willing to save one-half of his resources in order to provide for second-period consumption equal to first-period consumption even though he might die before the second period. On the other hand, in the limit as $\sigma\to 0$, the desire to smooth consumption is very weak so the consumer would consume all of his disposable resources in the first period.[9]

Using the consumption function (6), we can easily calculate the end-of-first-period wealth and the second-period consumption of the consumer. Combining equations (2) and (6) yields

$$(7)\quad W=(1-a)(B+Y-T)-aR^{-1}S-b.$$

Combining equations (3) and (6) yields the income expansion path

$$(8)\quad c_2=[(1-a)c_1-b]R/a,$$

which is linear and positively sloped.

II. Intergenerational Transfers

I have solved the consumer's saving-consumption decision conditional on the bequest B received at birth. In this section, I calculate the bequests received by each consumer. The bequest received by a consumer depends on the mortality history of the earlier generations of his family. Specifically, let j be the number of consecutive previous generations in a consumer's family that died at age 1 (i.e., did not live to the second period of their life). For example, $j=0$ indicates that the consumer's parent lived two periods and therefore left no bequest to the consumer. If $j=1$, then the consumer's parent died at age 1 leaving a bequest but the consumer's grandparent lived two periods leaving no bequest. All consumers are indexed according to j and I use the superscript j written in parentheses to indicate that a variable pertains to a consumer of type j. Observe that $p^j(1-p)$ is the fraction of consumers who are of type j.

First consider type-0 consumers. As indicated above, the parents of these consumers lived two periods, leaving no bequest so that $B^{(0)}=0$. The first-period consumption and end-of-first-period wealth of these consumers follow immediately from (6) and (7), respectively:[10]

$$(9)\quad c_1^{(0)}=a(Y-T+R^{-1}S)+b,$$

$$(10)\quad W^{(0)}=(1-a)(Y-T)-aR^{-1}S-b.$$

[9] If the consumer cannot borrow against his (uncertain) future Social Security benefit S, then $c_1\to B+Y-T$ as $\sigma\to 0$. Of course, if $S=0$, then the consumer will indeed consume all of his disposable lifetime resources in the first period.

[10] I assume that S and T are small enough and that the utility function and labor income are such that $W^{(0)}>0$. Note that if $b=0$ (as it would be with constant relative risk aversion), then $W^{(0)}>0$ provided that S and T are small enough.

Now consider consumers who receive positive bequests at birth, that is, consumers of type j, $j \geq 1$. Because all consumers have the same constant marginal propensity to consume, the difference in first-period consumption between any two consumers is proportional to the difference in the bequests they received at birth. In particular, the first-period consumption of a type-j consumer exceeds the first-period consumption of a type-0 consumer by $aB^{(j)}$:

$$c_1^{(j)} = aB^{(j)} + c_1^{(0)}. \tag{11}$$

Similarly, intracohort differences in end-of-first-period wealth are proportional to intracohort differences in the bequest received at birth, so that from (7) and (10) we obtain

$$W^{(j)} = (1-a)B^{(j)} + W^{(0)}. \tag{12}$$

Having related first-period consumption and end-of-first-period wealth to $B^{(j)}$, the bequest received at birth, the next step is to calculate $B^{(j)}$. If a type-$j-1$ consumer dies after one period, he leaves a bequest of $G^{-1}W^{(j-1)}$ to each of his children (who are type-j consumers). The bequest earns a gross rate of return R so that

$$B^{(j)} = (R/G)W^{(j-1)} \qquad j = 1,2,3,\ldots. \tag{13}$$

Substituting (13) into (12) yields the first-order linear constant coefficient difference equation $W^{(j)} = (1-a)(R/G)W^{(j-1)} + W^{(0)}$, $j = 1,2,3,\ldots$, which has the solution

$$W^{(j)} = W^{(0)} \sum_{i=0}^{j} (1-a)^i (R/G)^i \qquad j = 0,1,2,\ldots. \tag{14}$$

According to (14), as we increase the number of previous generations that died early leaving bequests, we increase $W^{(j)}$. We will assume that $(1-a)R < G$; hence, as j increases, $W^{(j)}$ approaches $W^{(0)}/[1-(1-a)(R/G)]$ asymptotically.[11]

I have now obtained a complete formal solution of the model. Given any nonnegative integer j, we know that a fraction $(1-p)p^j$ of the population is of type j. Then, using equations (9)–(11), (13), and (14), it is a simple matter to calculate the consumption, wealth, and bequests received at birth by each type-j consumer. The next step is to summarize the distributions of consumption, wealth and bequests by calculating the values of aggregate first-period consumption C_1^*, aggregate second-period consumption C_2^*, aggregate private wealth W^*, and aggregate bequests B^*. Each of these aggregates is expressed on a per capita basis (more precisely, per person in the young generation).[12]

Calculating the aggregate per capita values of both sides of (12), we obtain

$$W^* = (1-a)B^* + W^{(0)}. \tag{15}$$

Because a fraction p of each type of consumer dies early leaving a bequest, aggregate wealth held by consumers who die young is pW^*. Including the accrued interest on this wealth and adjusting for the fact that each generation has G times as many consumers as the previous generation, we obtain

$$B^* = p(R/G)W^*. \tag{16}$$

Substituting (16) into (15) yields

$$W^* = W^{(0)}/(1-(1-a)pR/G). \tag{17}$$

Therefore, per capita wealth is proportional to $W^{(0)}$, the wealth of type-0 consumers, and the constant of proportionality is independent of the tax parameters T and S.

[11]Since $0 < a < 1$, it follows immediately that if $R \leq G$, then the convergence condition $(1-a)R < G$ holds. To examine the case where $R > G$, observe from (6b) that $(1-a)R = \phi R^{1/(1-\gamma)}/[1+\phi R^{\gamma/(1-\gamma)}]$, where $\phi \equiv [(1-p)\delta]^{1/(1-\gamma)}$. Therefore, $(1-a)R < G$ if and only if $\phi R^{1/(1-\gamma)}[1-G/R] < G$. If $R > G$, then the convergence condition holds if and only if $\phi < [G/(R-G)]R^{-\gamma/(1-\gamma)}$.

[12]For example, aggregate private wealth per capita is defined as $W^* = \sum_{j=0}^{\infty}(1-p)p^jW^{(j)}$. Since only a fraction $(1-p)$ of young consumers survives to the second period of life, and since each generation is only G^{-1} times as large as the succeeding generation, aggregate second-period consumption is

$$C_2^* = (1-p)G^{-1}\sum_{j=0}^{\infty}(1-p)p^jc_2^{(j)}.$$

Aggregate economywide private consumption per capita, $C_1^* + C_2^*$, is equal to the sum of after-tax labor income, $Y-T$, plus Social Security payments to the surviving fraction $(1-p)$ of the old cohort, plus the net return on wealth, adjusted for population growth,[13]

$$(18)\quad C_1^* + C_2^* = Y - T + (1-p)G^{-1}S + (R/G - 1)W^*.$$

A final useful relation between C_1^* and C_2^* is obtained by calculating the per capita values of both sides of the income expansion path in (8) and recalling that the old cohort has $(1-p)G^{-1}$ times as many consumers as the young cohort,

$$(19)\quad C_2^* = (1-p)G^{-1}\left[(1-a)C_1^* - b\right]R/a.$$

Thus, C_1^* and C_2^* move in the same direction in response to changes in labor income Y, or in the Social Security parameters S and T.

III. The Effects of Actuarially Fair Social Security

In this section I consider the effects on savings and consumption of the introduction of a fully funded actuarially fair Social Security system. Let us suppose that the only role of the government is to collect Social Security taxes from the young and distribute Social Security benefits to the old. Thus the taxes T levied on the young are Social Security taxes. An actuarially fair Social Security system would levy a tax of $(1-p)R^{-1}$ dollars for each dollar of expected benefits, that is, $RT = (1-p)S$. Under this system, a young consumer contributes $(1-p)R^{-1}S$ to the Social Security system. He receives S if he survives to the second period of life, but receives zero if he dies after one period. Thus the expected present value of the Social Security benefit is $(1-p)R^{-1}S$ which is equal to the consumer's contribution. Put differently, the Social Security system runs a balanced account vis-à-vis each generation. The Social Security system collects taxes from the members of each generation when they are young, invests the tax revenue at a gross rate of return R, and then returns all of the tax revenue with accrued interest to the surviving old members of the generation.

A. *Aggregate Consumption and Capital Accumulation*

In order to study the effects of actuarially fair Social Security on aggregate consumption, I proceed in three steps. First, I analyze the effects of Social Security on the saving and consumption behavior of type-0 consumers. Then, I use the results about the effects on $W^{(0)}$ to analyze the effects on the private capital stock and on the total national capital stock. Finally, the relations between the national capital stock and aggregate consumption are used to determine the effects on C_1^* and C_2^*.

To calculate the effects of actuarially fair social security on consumption and saving of young type-0 consumers, we substitute $T = (1-p)R^{-1}S$ into (9) and (10) to obtain

$$(20)\quad c_1^{(0)} = aY + b + apR^{-1}S,$$

$$(21)\quad W^{(0)} = (1-a)Y - b - T - apR^{-1}S.$$

The introduction of actuarially fair Social Security increases the present value of lifetime resources, $B + Y - T + R^{-1}S$, by $-T + R^{-1}S = pR^{-1}S$. A consumer who survives to the second period receives a Social Security payment S that exceeds the value of his contribution with accrued interest, RT, because the surviving members of each generation receive (on a pro rata basis) the taxes-cum-interest contributed by members of their generation who died after one period. The effect of this increase in lifetime resources is to increase $c_1^{(0)}$ by $apR^{-1}S$.[14] The wealth

[13] Observe from (2) and (16) that $C_1^* = B^* + Y - T - W^* = Y - T - (1 - pR/G)W^*$. Since $c_2^{(j)} = RW^{(j)} + S$, and since the old cohort is $(1-p)G^{-1}$ times as large as the young cohort, we obtain $C_2^* = (1-p)G^{-1}(RW^* + S)$. Adding together the expressions for C_1^* and C_2^* yields (18).

[14] An alternative explanation for the increase in consumption is that a consumer's claim to Social Security

held at the end of the first period by type-0 consumers is reduced for two reasons: first, disposable resources available in the first period fall by the amount of the tax T; second, the increase in first-period consumption further reduces wealth held at the end of the first period.[15]

In a fully funded Social Security system, the total national capital stock per capita (measured at the end of a period) is equal to the sum of the aggregate private capital stock per capita, W^*, and the per capita capital stock held by the Social Security system T. Recall from (17), that the private capital stock W^* is proportional to $W^{(0)}$, and that the constant of proportionality does not depend on the parameters of the Social Security system. Since, from (21), the introduction of Social Security reduces $W^{(0)}$, it also reduces the aggregate private capital stock. Since $B^* = (pR/G)W^*$, the reduction in the aggregate private capital stock implies an equiproportionate reduction in aggregate bequests per capita.

The effect of actuarially fair Social Security on the aggregate national capital stock per capita, $W^* + T$, is easily determined by first observing from the definition of end-of-first-period wealth in (2) that

$$W^* + T = Y + B^* - C_1^*. \tag{22}$$

Then calculating the aggregate per capita values of both sides of (11) we obtain

$$C_1^* = aB^* + c_1^{(0)}. \tag{23}$$

Substituting (23) into (22) yields

$$W^* + T = Y + (1-a)B^* - c_1^{(0)}. \tag{24}$$

Since the introduction of Social Security causes B^* to fall and $c_1^{(0)}$ to increase, it is clear from (24) that the aggregate national capital stock $W^* + T$ is reduced by the introduction of Social Security.

Next I examine the effects on aggregate consumption of the introduction of actuarially fair Social Security. Substituting RT for $(1-p)S$ in (18) gives

$$C_1^* + C_2^* = Y + ((R/G)-1)(W^* + T). \tag{25}$$

Thus, aggregate private consumption is equal to the sum of labor income Y and the net return (adjusted for population growth) on national wealth. Observe that if $R = G$, so that the net rate of return on capital is equal to the rate of population growth, then the coefficient on national wealth in (25) is zero. In this case, $C_1^* + C_2^*$ is independent of the level of actuarially fair Social Security taxes and benefits. Furthermore, in view of the aggregate income expansion path in (19), both C_1^* and C_2^* are independent of the level of actuarially fair Social Security taxes and benefits when $R = G$. If $R > G$, then the reduction in aggregate wealth, $W^* + T$, induced by the introduction of Social Security, leads to a reduction in $C_1^* + C_2^*$; in light of (19), C_1^* and C_2^* are each reduced by the introduction of actuarially fair Social Security. Finally, if $R < G$, then C_1^* and C_2^* are each increased by the introduction of actuarially fair Social Security.[16]

benefits can be viewed as an annuity. If the consumer survives until retirement, the annuity pays some specified amount, but if the consumer dies before retirement, the annuity pays zero. Under an actuarially fair Social Security system, the price that the consumer pays for this annuity (i.e., the Social Security tax levied on young consumers) is equal to the expected present value of future payoffs. However, consumers would be willing to pay more than the expected present value of future payoffs because the payoffs are positively correlated with future marginal utility of consumption. The annuity has a positive payoff if and only if the consumer survives, thereby having a positive marginal utility of consumption; the annuity has a zero payoff if the consumer dies, in which case wealth has zero marginal utility. Therefore, an actuarially fair increase in the level of Social Security taxes and benefits will make a young consumer wealthier and hence increase his consumption.

[15] In a balanced budget pay-as-you-go system, $GT = (1-p)S$. In this case, equation (9) implies $c_1^{(0)} = aY + b + a((G/((1-p)R))-1)T$ so that the introduction of Social Security causes $c_1^{(0)}$ to increase, decrease, or remain unchanged according to whether G is greater than, less than, or equal to $(1-p)R$. It follows from equation (10) that the introduction of Social Security causes $W^{(0)}$ to fall by $(1-a)T + aR^{-1}S$.

[16] Under a balanced budget pay-as-you-go system, $GT = (1-p)S$, so that from (18) aggregate consumption is $C_1^* + C_2^* = Y + (R/G - 1)W^*$. As shown in fn.

B. *The Intracohort Distributions of Consumption and Wealth*

Having analyzed the effects of Social Security on the aggregate consumption of the young cohort and the aggregate consumption of the old cohort, I now examine the intracohort distributions of consumption and wealth. As already shown (equation (20)), the first-period consumption of type-0 consumers increases by $apR^{-1}S$ in response to the introduction of Social Security. Also I have shown that $W^{(0)}$ falls by $T + apR^{-1}S$ when Social Security is introduced. As a consequence of the fall in $W^{(0)}$, there is a reduction in bequests, $B^{(1)}$, received at birth by type-1 consumers. Indeed, the introduction of Social Security reduces $B^{(j)}$ for all type-j consumers for $j = 1, 2, 3, \ldots$. This result follows from the facts that $B^{(j)}$ is proportional to $W^{(0)}$ (see equations (13) and (14)) and that $W^{(0)}$ is reduced by the introduction of Social Security. Below I analyze the effects of the induced reduction in bequests in the intracohort distribution of consumption.

The deviation of a type-j consumer's first-period consumption from the average level of first-period consumption is proportional to $B^{(j)} - B^*$ (see equations (11) and (23)):

$$c_1^{(j)} - C_1^* = a(B^{(j)} - B^*). \tag{26}$$

Since $B^{(j)}$ and B^* are each proportional to $W^{(0)}$, it follows that $c_1^{(j)} - C_1^*$ is also proportional to $W^{(0)}$.[17] Because the introduction of actuarially fair Social Security reduces $W^{(0)}$, it also reduces the (magnitude of the) deviation of type-j consumer's first-period consumption from the average first-period consumption.[18] Thus, the distribution of consumption is narrowed by the introduction of Social Security. More precisely, all central moments of the intracohort distribution of $c_1^{(j)}$ are reduced by the introduction of Social Security.

The effects of the introduction of Social Security on second-period consumption are easily calculated by observing from (8) that $c_2^{(j)}$ can be expressed as an increasing linear function of $c_1^{(j)}$. Therefore, the narrowing of the distribution of $c_1^{(j)}$ implies that the distribution of $c_2^{(j)}$ is also narrowed by the introduction of Social Security.

For the case in which $R = G$, it is straightforward to analyze the (steady-state) welfare implications of the introduction of Social Security. In this case, the introduction of actuarially fair Social Security does not affect the average levels of consumption of the young or of the old as explained in Section III, Part A. However, it narrows the distribution of consumption of each cohort. Therefore, if each consumer has an identical utility function and receives equal weight in the social welfare function, the introduction of Social Security is welfare improving. If $R < G$, then the introduction of Social Security raises the average level of consumption and reduces the variance of consumption. Each of these effects increases social welfare. However, if $R > G$, then the introduction of Social Security reduces average consumption, which tends to reduce welfare, but also reduces the intracohort variance of consumption, which tends to raise welfare.

IV. The Transition Path to the New Steady State

The analysis in Section III of the effects of the introduction of actuarially fair Social Security was a comparative steady-state anal-

15, $W^{(0)}$ is reduced by the introduction of pay-as-you-go Social Security. Therefore, since W^* is proportional to $W^{(0)}$, it follows that W^* is also reduced. As in the text, aggregate consumption is reduced, increased, or left unchanged according to whether R is greater than, less than, or equal to G.

[17]More formally, using equations (13), (14), (16), and (17), equation (26) can be rewritten as

$$c_1^{(j)} - C_1^* = a(R/G)W^{(0)}$$
$$\times\left(\sum_{i=0}^{j-1}(1-a)^i(R/G)^i - \frac{p}{1-p(R/G)(1-a)}\right)$$

where $\sum_{i=0}^{j-1}(1-a)^i(R/G)^i$ is equal to zero for $j = 0$.

[18]Recall from fn. 15 that $W^{(0)}$ is also reduced by pay-as-you-go Social Security. Hence, pay-as-you-go Social Security also narrows the distribution of consumption.

ysis; it was assumed that the Social Security system had been in effect long enough so that essentially no one received a bequest that included part of the savings of an ancestor who lived in the initial regime without Social Security. Equivalently, it was assumed that each person had at least one ancestor who lived for two periods under the new regime, leaving no bequests and thus severing links to the old regime.

In this section, I examine the transition path to the new steady state, which accompanies the introduction of an actuarially fair Social Security system. I show that the introduction of Social Security reduces the intracohort variances of first-period and second-period consumption for every generation (except the first) born under the new Social Security regime. Also, if $R \leq G$, then the average levels of first-period and second-period consumption of each generation are at least as high under the Social Security regime as in the absence of Social Security. In this case, the introduction of Social Security increases the welfare of every generation born under the Social Security regime.

Suppose that actuarially fair fully funded Social Security is introduced at the beginning of period t^*+1. It will be assumed that since the older cohort (born at time t^*) did not contribute to the Social Security system, they receive no benefits. The young generation (born at time t^*+1) pays a tax $T=(1-p)R^{-1}S$ and the survivors will each receive a Social Security payment of S as discussed in Section III. The bequests received by each individual in the young generation are invariant to the introduction of Social Security, and for a given level of bequests received at birth, the introduction of Social Security increases the present value of lifetime income by $pR^{-1}S$. Thus, every consumer in this generation increases first-period consumption by $apR^{-1}S$, and every survivor increases second-period consumption by $(1-a)pS$. This generation unanimously favors the introduction of actuarially fair Social Security.

Next I consider the effect of the introduction of Social Security on subsequent generations. Let the subscript m denote that a variable pertains to a consumer born at the beginning of period t^*+m. Let Δ denote the change in a variable induced by the introduction of Social Security (relative to the regime without Social Security). Thus, the effect of Social Security on the first-period consumption of type-j consumers born at time t^*+m is obtained from (11) and (13) as

$$\Delta c_{1,m}^{(j)} = \Delta c_{1,m}^{(0)} + a(R/G)\Delta W_{m-1}^{(j-1)}, \quad j=1,2,3,\ldots \quad m=2,3,4,\ldots. \tag{27}$$

Equation (27) displays the two countervailing effects on the consumption of subsequent generations. It follows immediately from (20) that

$$\Delta c_{1,m}^{(0)} = apR^{-1}S > 0. \tag{28}$$

Thus, as explained earlier, the first-period consumption of type-0 consumers increases.

For consumers who receive positive bequests at birth, there is a second effect on consumption and lifetime income because these consumers receive smaller bequests as a result of the introduction of Social Security. A straightforward generalization of (14) yields

$$\Delta W_{m-1}^{(j-1)} = \Delta W^{(0)} \sum_{i=0}^{j^*-1} (1-a)^i (R/G)^i \quad j=1,2,3,\ldots \quad m=2,3,4,\ldots. \tag{29}$$

where $j^* = \min(j, m-1)$. Observe from (21) that[19]

$$\Delta W^{(0)} = -\left(T + apR^{-1}S\right) < 0. \tag{30}$$

Since $\Delta W^{(0)} < 0$, it is clear from (29) that $\Delta W_{m-1}^{(j-1)} < 0$ so that type-j consumers born at the beginning of period t^*+m receive smaller bequests at birth. The magnitude of the reduction in bequests is strictly increasing in j^*. Thus, for the generation born at

[19] Recall from fn. 15 that $\Delta W^{(0)} < 0$ for pay-as-you-go Social Security also.

the beginning of period t^*+m, the *reduction* in bequests received by type-j consumers is strictly increasing in j for $j=0,1\ldots, m-1$, and is constant for $j=m-1, m, m+1,\ldots$. This finding combined with the fact that the *level* of bequests received by type-j consumers is strictly increasing in j for the Social Security regime as well as the regime without Social Security implies that the introduction of Social Security reduces the intracohort variance of bequests received by all generations born after period t^*+1.[20] Since first-period (second-period) consumption is a linear function of the bequest received at birth, the introduction of Social Security also reduces the intracohort variance of first-period (second-period) consumption for these generations.

I have derived unambiguous results about the intracohort variance of consumption along the transition path to the new steady state. The effects on the average level of consumption are less clear-cut. As already shown, for the generation born at the beginning of period t^*+1, the average levels of first-period and second-period consumption are increased by the introduction of Social Security. It has also been shown that, in the new steady state, the average levels of $c_1^{(j)}$ and $c_2^{(j)}$ decrease, increase, or remain unchanged, depending on whether R is greater than, less than, or equal to G. I show in the Appendix that $\Delta C^*_{1,m}$, the change in the average level of first-period consumption of the generation born at time t^*+m, is

$$(31)\quad \Delta C^*_{1,m}=\frac{apR^{-1}S}{1-p(R/G)(1-a)} \times\{1-(R/G)+[1-p(1-a)] \times(R/G)^m p^{m-1}(1-a)^{m-1}\}.$$

Since it has been assumed that $(1-a)pR$ is less than G, it is clear from (31) that $\Delta C^*_{1,m}$ decreases as m increases. The reason is that as m increases (i.e., as we increase the length of time for which the Social Security regime has been in effect), there is a decrease in the amount of bequests which represent accumulated saving from generations born before the introduction of Social Security, when private saving was higher.

In the case in which $R=G$, equation (31) implies that $\Delta C^*_{1,m}$ is equal to $aR^{-1}Sp^m(1-a)^{m-1}$, which is positive for all finite m. Thus, since the introduction of Social Security increases the average and reduces the variance of $c_{1,m}^{(j)}$ for all finite m, it also (see equation (8)) increases the average value and reduces the variance of $c_{2,m}^{(j)}$ for all finite m. Therefore, if $R=G$, the introduction of Social Security is welfare improving for every generation born under the new Social Security regime. More generally, if $R\leq G$, the welfare of every generation (except the current old generation which is unaffected) is improved by the introduction of Social Security.

The welfare effects of the introduction of Social Security are less clear-cut in the case in which $R>G$. Clearly, the welfare of the generation born at time t^*+1 is improved because, as explained earlier, the first-period consumption of every consumer in this generation increases by $apR^{-1}S$ (and from equation (8), second-period consumption increases by $(1-a)pS$). For all generations born after time t^*+1, the introduction of Social Security reduces the intracohort variance of consumption. For sufficiently small m, it follows from (31) that the average level of first-period (and second-period) consumption is increased by the introduction of Social Security. Thus, for these generations, welfare is increased. The difficulty in my welfare analysis arises for generations born long after the introduction of Social Security. If $R>G$, then it follows from (31) that for sufficiently large m, the average first-period (and a fortiori average second-period) consumption of the generation born at time t^*+m is reduced by the introduction of Social Security. The effect on the welfare of this generation thus depends on whether the welfare-improving effects of reduced variance dominate the welfare-worsening effects of reduced average consumption.

[20] This statement is simply an application of the fact that if $x(j)$ and $y(j)$ are strictly increasing in j, and if $y(j)-x(j)$ is nonincreasing in j and strictly decreasing for some j, then the variance $y(j)$ is less than the variance of $x(j)$.

V. Private Annuities

In previous sections in this paper, it was assumed that there is no private market for annuities, and I showed that the introduction of actuarially fair Social Security reduces the national capital stock. However, if there were a competitive market for annuities, then the introduction of Social Security would have no effect because the competitively supplied actuarially fair annuities would be perfect substitutes for actuarially fair Social Security; hence, consumers could completely offset the effects of Social Security by conducting transactions in the private annuity market. Since the introduction of actuarially fair Social Security reduces the steady-state capital stock, it is natural to ask whether the introduction of a competitive annuity market also reduces the steady-state capital stock. In this section, we analyze a simple example to show that the introduction of a market for private annuities can either increase or decrease the steady-state capital stock.[21]

With the introduction of private annuities, there are now two alternative forms in which a consumer can hold his wealth. As before, he can hold capital directly, earning a gross rate of return R. Alternatively, he can deposit his savings at an annuity company. The annuity company operates by accepting deposits from young consumers and using these deposits to buy capital which earns a gross rate of return R. At the beginning of the following period, the annuity company distributes its holdings (with accumulated interest) to its surviving depositors in proportion to their initial deposits. Thus, each surviving depositor at the annuity company receives $A=R/(1-p)$ dollars for each dollar initially deposited. As shown by Yaari, consumers who do not have explicit bequest motives will choose to hold all of their wealth in the form of these annuities. Thus, there will be no bequests.

Consumers can, by holding annuities, earn a gross rate of return A on their savings so that $c_2=A[Y-c_1]$. The maximization problem of the representative consumer[22] is

$$\max_{c_1} U(c_1)+(1-p)\,\delta U(A(Y-c_1)). \tag{32}$$

The first-order condition for this problem is

$$U'(c_1)=(1-p)A\,\delta U'(c_2). \tag{33}$$

With actuarially fair annuities $(1-p)A=R$, so that the first-order condition (33) can be written as

$$U'(\hat{c}_1)=R\,\delta U'(\hat{c}_2), \tag{34}$$

where a circumflex denotes the value of a variable in the presence of a private annuity market.

For the remainder of this section we assume that $R=\delta=1$, that is, that the net rate of return on capital and the rate of time preference are each equal to zero. With $R\delta=1$, (34) implies that $\hat{c}_1=\hat{c}_2$ for any strictly concave utility function $U(\;)$. Since $\hat{c}_2=A(Y-\hat{c}_1)$ and $A=(1-p)^{-1}$, we obtain $\hat{c}_1=\hat{c}_2=(1/(2-p))Y$. Therefore, since $\hat{W}=Y-\hat{c}_1$, we obtain

$$\hat{W}=((1-p)/(2-p))Y. \tag{35}$$

Now consider the economy without an annuity market. For the remainder of this section, it will be assumed that $U(c)$ exhibits constant relative risk aversion. Recall from Section I that with $R=\delta=1$, and a constant coefficient of relative risk aversion equal to σ, the marginal (and average) propensity to consume is $a=[1+(1-p)^{1/\sigma}]^{-1}$. Thus, the first-period consumption of type-0 consumers is

$$c_1^{(0)}=Y/\left(1+(1-p)^{1/\sigma}\right). \tag{36}$$

It is straightforward to show that

$$\hat{c}_1 \gtreqless c_1^{(0)} \quad \text{as} \quad \sigma \gtreqless 1. \tag{37}$$

[21] See Kotlikoff, John Shoven, and Spivak (1983) for an analysis of the effects of various annuity arrangements on capital accumulation.

[22] Since there are no bequests, there is no need to distinguish consumers according to the mortality history of their families. Also, since actuarially fair Social Security has no effect in the presence of an annuity market, I simply set $S=T=0$.

The intuition for this result is that the introduction of a private annuity market raises the rate of return on private savings from R to $R/(1-p)$. The income effect of this change is to raise first-period consumption whereas the substitution effect is to reduce first-period consumption. With $\sigma > 1$, the income effect dominates and with $\sigma < 1$, the substitution effect dominates. For logarithmic utility, ($\sigma = 1$), the income and substitution effects exactly offset each other. By contrast, notice that although Social Security has the payoff characteristics of an annuity, the introduction of actuarially fair Social Security has a positive income effect but has no substitution effect because individual consumers cannot choose the level of savings to be held in the form of Social Security.

The analysis in the above paragraph examines the effect on type-0 consumers of the introduction of an annuity market. For type-j consumers, there is an additional effect, because these consumers receive bequests in the absence of private annuities but do not receive bequests in the presence of annuities. To calculate the effect of an annuity market on the long-run capital stock, observe from (10) and (17) that in the absence of annuities, and with $R = \delta = 1$, the steady-state capital stock is

$$W^* = Y/\left[1+(1-p)^{-1/\sigma} - p/G\right]. \tag{38}$$

Comparing (35) and (38) it can be shown that

$$\hat{W} \lesseqgtr W^* \quad \text{as} \quad \sigma \gtreqless \bar{\sigma} \tag{39}$$

where

$$\bar{\sigma} = \left[1 - \frac{\ln(1+((1-p)p/G))}{\ln(1-p)}\right]^{-1} < 1.$$

According to (39), there is a critical value of the coefficient of relative risk aversion that determines whether the long-run capital stock increases or decreases when a private annuity market is introduced. When $\sigma > 1$, the introduction of an annuity market raises first-period consumption of type-0 consumers and thus reduces their saving. In addition, the elimination of bequests received by type-j consumers for $j \geq 1$ also tends to reduce private wealth.[23] On the other hand, when $\sigma < 1$, the introduction of private annuities reduce the first-period consumption and increases the saving of type-0 consumers. Whether this wealth-increasing effect dominates the wealth-reducing effect of eliminating bequests depends on whether σ is less than $\bar{\sigma}$.

VI. Concluding Remarks

I have developed a general equilibrium model of precautionary saving and accidental bequests that is sufficiently rich to produce endogenous distributions of consumption, wealth, and bequests. The model is based on individual utility-maximizing behavior and yields decision rules for consumers that are linear and easily aggregated. After developing the model in Sections I and II, it was shown in Section III that, in the absence of a private annuity market, the introduction of actuarially fair Social Security crowds out private wealth by more than one for one, thereby reducing national wealth; in addition, it reduces all central moments of the distributions of consumption, wealth, and bequests. Section IV analyzes the transition path to the new steady state when Social Security is introduced. The immediate effect is for the average level of consumption by young consumers to increase and for the variance of their consumption to remain unchanged. However, both the mean and the variance of consumption by young consumers decreases continually as

[23] Kotlikoff, Shoven, and Spivak examine a more complex overlapping-generations model with uncertain lifetimes. They solve their model numerically, and in each of their numerical simulations, they assume that the constant coefficient of relative risk aversion is greater than one. Although they find that the introduction of a perfect annuity market reduces long-run aggregate wealth in each of their similations, the results in (39) suggest caution in applying this result to consumers with a coefficient of relative risk aversion sufficiently below one. Of course, since their model differs somewhat from the model presented herein, the critical value of the coefficient of relative risk aversion, if one exists, would probably differ from that in (39).

each subsequent generation is born. In Section V, I switched attention from publicly provided annuities to privately traded annuities and showed that the introduction of a private annuity market would cause the steady-state capital stock to increase or decrease depending on the risk aversion of consumers.

The model presented in this paper was purposely designed to allow a simple examination of precautionary saving and accidental bequests in a general equilibrium framework. Toward this end, the following simplifying assumptions were made: 1) consumers live for either one or two periods; 2) the rate of return on capital, R, is constant; 3) there is no private market for annuities (except in Section V); and 4) consumers are selfish, that is, they have no bequest motive.

An implication of assumption 1 and the assumption that consumers give birth to their children at the end of the first period of life is that each consumer knows at birth exactly what bequest he will receive from his parent. If consumers lived potentially for many periods so that a parent's lifetime uncertainty were not resolved when the child is born, then we would have the additional problem of calculating optimal consumption behavior when there is the prospect of receiving a bequest of uncertain size at an unknown date in the future. Edi Karni and Itzhak Zilcha (1984) have examined the case in which consumers live for three periods. They prove the existence of the steady-state equilibrium in the absence of annuity and insurance markets, and demonstrate that the introduction of competitive life insurance and annuity markets leads to a Pareto optimal steady-state equilibrium. However, they do not provide closed-form solutions for consumption. Also, their model cannot be used to examine long-run capital accumulation since capital is absent from their model. Kotlikoff, Shoven, and Spivak also relax assumption 1, but provide numerical rather analytic solutions of their model.

Assumptions 1 and 2 are both relaxed by Glenn Hubbard (1984). In place of the linear technology assumed above, Hubbard introduces a neoclassical production function into a model with uncertain lifetimes. However, he assumes that the government confiscates the assets held by consumers when they die and then redistributes the assets in lump sum fashion. This assumption circumvents the technical difficulty mentioned above but this simplification also eliminates the intracohort variations in consumption and wealth. Hubbard does not solve his model analytically and resorts to numerical simulation to study the effects of Social Security.

Assumption 3 is crucial in order for Social Security to have an effect in this model. If there were a competitive annuity market, the rate of return on competitively supplied annuities would be equal to the implicit rate of return offered by Social Security. In this case consumers could undo the effects of Social Security by conducting offsetting transactions in the private annuity market. However, if the probability of dying after one period of life differed across consumers, then Social Security would have an effect on behavior. In another paper (1984a), I assume that an individual's probability of an early death is private information known only by the individual, so that the private annuity market is subject to adverse selection. However, a compulsory Social Security system is immune to adverse selection and can offer a higher rate of return than the equilibrium rate of return in the private annuity market. Thus, consumers cannot effectively undo the effects of Social Security by transacting in the private annuity market because private and social annuities are no longer perfect substitutes.

Eckstein, Eichenbaum, and Peled (1985b) also use the insight that the Social Security system is immune to adverse selection. They assume that consumers have no bequest motive so that with a private annuity market, there are no bequests, accidental or otherwise. However, in my 1984a paper, I relax assumption 4 and specify the utility function to have a bequest motive. I then show that the introduction of Social Security can either increase or decrease the steady-state capital stock depending on the strength of the bequest motive. In my 1984b paper, I assume that the probability of an early death differs across consumers, but these probabilities are public information. In this case, the intro-

duction of actuarially fair Social Security will have an effect if the government chooses not to discriminate on the basis of the probability of dying.

Although I have made some progress in incorporating a bequest motive into an overlapping generations model with uncertain lifetimes, further research is needed. My other papers (1984a,b) used the Hakansson, Fischer, and Scott Richard (1975) utility function which specifies a consumer's utility as a function of his own consumption and of the size of the bequest he leaves. An alternative formulation is based on Barro's intergenerational altruism in which a consumer derives utility from his own consumption and from the utility of his heirs. This formulation effectively converts the individual consumer's decision problem into an infinite horizon problem. In future research I plan to study the role for fiscal policy in an overlapping generations economy populated by consumers with uncertain lifetimes and altruistic bequest motives.

Appendix

Here I derive equation (31) in the text which shows the effect on $C^*_{1,m}$ of the introduction in period t^*+1 of actuarially fair Social Security.

It will be useful to define x as

$$\text{(A1)} \qquad x \equiv (1-a)(R/G).$$

Under actuarially fair Social Security, $RT = (1-p)S$ so that

$$\text{(A2)} \qquad (R/G)(T + apR^{-1}S) = (R/G)(1-p+ap)R^{-1}S = ((R/G)-px)R^{-1}S.$$

Substituting (28)–(30) into (27) and using (A1) and (A2), we obtain

$$\text{(A3)} \qquad \Delta c^{(j)}_{1,m} = aR^{-1}S\left\{p-[(R/G)-px]\sum_{i=0}^{j^*-1}x^i\right\}.$$

As a step toward calculating the average value of each side of (A3), I first calculate

$$\text{(A4)} \qquad \sum_{j=0}^{\infty}(1-p)p^j\sum_{i=0}^{j^*-1}x^i = (1-p)\sum_{j=0}^{\infty}p^j\frac{1-x^{j*}}{1-x}.$$

Recalling that $j^* = \min(j, m-1)$, (A4) can be rearranged to yield

$$\text{(A5)} \qquad \sum_{j=0}^{\infty}(1-p)p^j\sum_{i=0}^{j^*-1}x^i = \frac{1-p}{1-x}\left[\frac{1}{1-p}-\sum_{j=0}^{m-1}p^jx^j-\sum_{j=m}^{\infty}p^jx^{m-1}\right].$$

Calculating the sums on the right-hand side of (A5) yields

$$\text{(A6)} \qquad \sum_{j=0}^{\infty}(1-p)p^j\sum_{i=0}^{j^*-1}x^i = \frac{1-p}{1-x}\left[\frac{1-p^mx^{m-1}}{1-p}-\frac{1-p^mx^m}{1-px}\right],$$

which can be simplified to yield

$$\text{(A7)} \qquad \sum_{j=0}^{\infty}(1-p)p^j\sum_{i=0}^{j^*-1}x^i = (p/(1-px))(1-p^{m-1}x^{m-1}).$$

Now calculate the average value of each side of (A3) and use (A7) to obtain

$$\text{(A8)} \qquad \Delta C^*_{1,m} = apR^{-1}S\left\{1-\frac{(R/G)-px}{1-px}\times(1-p^{m-1}x^{m-1})\right\}.$$

Rearranging (A8) yields

$$\text{(A9)} \qquad \Delta C^*_{1,m} = \frac{apR^{-1}S}{1-px}\{1-(R/G)+[(R/G)-px]p^{m-1}x^{m-1}\}.$$

Recognizing that

$$\text{(A10)}\quad [(R/G)-px]\,p^{m-1}x^{m-1} = [1-p(1-a)]\,p^{m-1}(1-a)^{m-1}(R/G)^{m}$$

then yields equation (31) in the text.

REFERENCES

Abel, Andrew B., (1984a) "Capital Accumulation and Uncertain Lifetimes with Adverse Selection," mimeo., Harvard University, 1984.

——, (1984b) "The Effects of Social Security in the Presence of Perfect Annuity Markets," mimeo., Harvard University, 1984.

Barro, Robert J., "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November/December 1974, *82*, 1095–1117.

—— **and Friedman, James W.**, "On Uncertain Lifetimes," *Journal of Political Economy*, August 1977, *85*, 843–49.

Bernheim, B. Douglas, Shleifer, Andrei and Summers, Lawrence, "Bequests as a Means of Payment," Working Paper No. 1303, National Bureau of Economic Research, March 1984.

Diamond, Peter, "National Debt in a Neoclassical Growth Model," *American Economic Review*, December 1965, *55*, 1126–50.

Drazen, Allan, "Government Debt, Human Capital and Bequests in a Life-Cycle Model," *Journal of Political Economy*, June 1978, *86*, 505–16.

Eckstein, Zvi, Eichenbaum, Martin S. and Peled, Dan, (1985a) "The Distribution of Wealth and Welfare in the Presence of Incomplete Annuity Markets," *Quarterly Journal of Economics*, forthcoming 1985.

——, ——, **and** ——, (1985b) "Uncertain Lifetimes and the Welfare Enhancing Properties of Annuity Markets and Social Security," *Journal of Public Economics*, 1985 forthcoming.

Fischer, Stanley, "A Life Cycle Model of Life Insurance Purchases," *International Economic Review*, February 1973, *14*, 132–52.

Hakansson, Nils, "Optimal Consumption and Investment Strategies Under Risk, An Uncertain Lifetime and Insurance," *International Economic Review*, October 1969, *10*, 443–66.

Hubbard, R. Glenn, "'Precautionary' Saving Revisited: Social Security, Individual Welfare and the Capital Stock," mimeo., Northwestern University, May 1984.

Karni, Edi and Zilcha, Itzhak, "A Welfare Analysis of Steady States in an Overlapping Generations Model with Uncertain Lifetimes," Working Papers in Economics No. 137, Department of Political Economy, Johns Hopkins University, April 1984.

Kotlikoff, Laurence J. and Spivak, Avia, "The Family as an Incomplete Annuities Market," *Journal of Political Economy*, April 1981, *89*, 372–91.

——, **Shoven, John and Spivak, Avia**, "Annuity Markets, Savings, and the Capital Stock," Working Paper No. 1250, National Bureau of Economics, December 1983.

—— **and Summers, Lawrence**, "The Role of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy*, August 1981, *90*, 706–32.

Levhari, David and Mirman, Leonard, "Savings and Consumption with an Uncertain Horizon," *Journal of Political Economy*, April 1977, *85*, 265–81.

Modigliani, Franco and Brumberg, Richard, "Utility Analysis and Aggregate Consumption Functions: An Attempt at Integration," in Andrew Abel, ed., *The Collected Papers of Franco Modigliani*, Vol. II, Cambridge: MIT Press, 1979.

Richard, Scott F., "Optimal Consumption, Portfolio and Life Insurance Rules for an Uncertain Lived Individual in a Continuous Time Model," *Journal of Financial Economics*, June 1975, *2*, 187–203.

Samuelson, Paul A., "An Exact Consumption-Loan Model of Interest with or Without the Social Contrivance of Money," *Journal of Political Economy, December* 1958, *66*, 467–82.

Sheshinski, Eytan and Weiss, Yoram, "Uncertainty and Optimal Social Security Systems," *Quarterly Journal of Economics*, May 1981, *96*, 189–206.

Yaari, Menachem E., "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer," *Review of Economics Studies*, April 1965, *32*, 137–50.

A Test of Dual Labor Market Theory

By WILLIAM T. DICKENS AND KEVIN LANG*

This paper presents a test of two of the most important claims of dual market theory—that there is a distinct low-wage (secondary) labor market in which there are no returns to schooling and workers do not receive on-the-job training, and that there are noneconomic barriers that prevent at least some secondary workers from obtaining better (primary) jobs.

Human capital theory has tended to emphasize differences among people, rather than among jobs, as a determinant of the distribution of income. Workers in low-wage jobs are viewed simply as low-productivity workers who are unwilling or unable to obtain the skills that are necessary for access to higher paying jobs. It follows from this approach that the way to eliminate poverty is to provide individuals with more skills, or with incentives to obtain skills.

Dual market theorists have maintained that jobs can be roughly divided into two groups: those with low wages, bad working conditions, unstable employment, and little opportunity for advancement (secondary jobs), and those with relatively high wages, good working conditions and opportunities for advancement into higher paying jobs (primary jobs) (Peter Doeringer and Michael Piore, 1971). Advocates of this view have argued that primary sector jobs are rationed, and that, in particular, women, blacks, and other minorities find it difficult to obtain primary employment. Since, in the view of dual market theorists (Suzanne Berger and Piore, 1980), it is unlikely that rationing can be eliminated, training programs will not be successful in eliminating poverty and the major roles for policy are providing income support, ensuring that the rationing system is "fair," and minimizing the extent of the secondary sector by stabilizing aggregate demand.

Despite significant differences in their views of the low-wage labor market, neither the advocates of dual market theory nor its critics have specified potentially conclusive tests of either the dual market typology or the hypothesis of noneconomic barriers to entering the primary sector. Difficulties arise because tests of the dual market hypothesis often rely on circular definitions of the sectors.

We propose strong tests of both hypotheses. Our results provide considerable support for the view that there are two distinct labor markets—a primary labor market with a wage profile similar to that predicted by human capital theory, and a secondary market with a completely flat (low) wage profile. Our results also provide support for the hypothesis that there are noneconomic barriers that prevent nonwhites from entering the primary sector.

In Section I, we review some of the most noteworthy empirical work on dual market theory. In the second section, we outline what we consider to be the essential differences between dual market and human capital theory, and develop a formal test that allows us to distinguish between the two hypotheses. The results are presented in Section III.

I. A Partial Review of Empirical Work on Dual Market Theory

Although advocates of dual market theory may differ on the particulars, all agree on two basic tenets:

1. The dual market typology described above is a useful characterization—most jobs

*Department of Economics, University of California, Berkeley, CA 94720 and National Bureau of Economic Research, and School of Social Sciences, University of California, Irvine, CA 92717, respectively. We acknowledge the able research assistance of Kathy Marshall and Ed Plummer, the generous support of the Institute of Industrial Relations at Berkeley and the National Science Foundation under grant no. SES-8409380, and helpful advice from Paul Ruud, Tom Rothenberg, James Albrecht, and Steve Woodbury. This work has also benefited from comments at workshops at UC-Berkeley and Irvine, University of Southern California, and the NBER Labor Studies Group. The opinions expressed are our own. No responsibility should be ascribed to any other group or individual.

strongly resemble the description of either primary or secondary jobs.

2. At most times there is rationing of primary sector jobs.

A number of attempts have been made to test either or both of these hypotheses.

Studies of the validity of the dual market typology have taken two forms, factor analysis of job and/or worker characteristics, and comparisons of wage equations for different occupations and industries. The authors who have used factor analysis have found a dominant factor fitting the dual market typology and have found bimodal distributions of factor scores (David Gordon, 1971; Robert Buchele, 1976a,b; Gerry Oster, 1979). However, the correlation of certain attributes such as low wages and bad working conditions does not provide strong support for the dual market hypothesis of the existence of sectors with distinct wage-setting mechanisms.

Consequently, some researchers have attempted to test more directly the hypothesis that the wage-setting mechanisms are different in the two sectors. The approach these authors have followed is to divide occupations and/or industries into two sectors on the basis of the characteristics of the jobs, or of workers in those occupations or industries. Having thus divided the sample, they test for differences in the wage equations for the two sectors. Some have found patterns corresponding roughly to dual market theory (Paul Osterman, 1975; M. Carnoy and R. Rumberger, 1980; Buchele, 1976a,b; Samuel Rosenberg, 1976; Eric Wright, 1979); others have found little support for the hypothesis (Lynne Zucker and Carol Rosenstein, 1981; Robert Bibb and William Form, 1977; Randy Hodson, 1977). In addition, none of these studies has been entirely free of anomalies.

Unfortunately, dividing the sample on the basis of occupation or industry has major drawbacks. Since a worker's choice of industry or occupation is not independent of unmeasured characteristics, there is danger of sample selection bias. Often industries and occupations are classified as secondary because they offer low wages. It is not surprising to find that in low-wage jobs the return to schooling is relatively low (Glenn Cain, 1976). In addition, the assumption that all members of an occupation or industry are in the secondary sector may significantly reduce the power of the test. For example, no one would argue that managers and skilled workers in industries which employ a substantial number of secondary workers are themselves in secondary jobs. It is possible that the anomalous results found in this literature are due to inaccurate classification.

Both the factor analyses and attempts to test for the existence of distinct wage equations for the primary and secondary sectors described above are essentially concerned only with the dual market typology. As noted in the introduction, dual market theorists maintain not only that they have developed an accurate typology but that primary jobs are rationed. In fact, it is the latter position which constitutes the major break with human capital theory.

Several authors have suggested that the existence of distinct wage equations for the primary and secondary sectors would constitute a refutation of human capital theory (Buchele, 1976a,b; Osterman), but this is not the case. If an individual can move out of the secondary sector in order to obtain returns on experience or education, the existence of a sector in which there are no returns is inconsequential (Cain). Thus the basis of the allocation of workers between the sectors is crucial; are primary sector jobs rationed?

Several authors have addressed the issue of mobility between the two sectors. Duane Leigh (1976) finds substantial and comparable earnings growth for black and white workers and suggests that this refutes the dual market hypothesis. Bradley Schiller (1977) reports extensive upward mobility of individuals at the bottom of the income distribution during the period 1957 to 1971. He argues that this constitutes a refutation of dual market theory.

On the other hand, Rosenberg (1976) and Carnoy and Rumberger find that minority workers are more likely to begin their careers in the secondary sector and, having started there, are less likely to leave than are whites. Rosenberg also finds that human capital variables do not help to explain the upward mobility of minority workers. These authors argue that this differential mobility supports dual market theory.

In fact, measuring mobility does not provide a test of rationing of primary market jobs. As Rosenberg (1979) notes, some mobility is consistent with dual market theory, while purely random movement is not implied by human capital theory. It is easy to derive a simple human capital model with firm-specific training in which there is no mobility between jobs whatsoever. No one has specified, and it is probably impossible to do so correctly, what levels of mobility would constitute refutations of dual market or human capital theory. Although studies of differential mobility between races are suggestive, the key issue is whether there are qualified individuals who would like to work in the primary sector but cannot find a job there. No study has addressed this issue.

Thus empirical work contrasting dual market and human capital theory has suffered from two major drawbacks. The taxonomies that have been developed simultaneously bias the results in favor of the dual market hypothesis by virtue of the selection criteria and are too gross to allow accurate testing of the hypothesis. Furthermore, the crucial issue of barriers to entry has not been addressed.

In Section II we propose a technique that allows us to derive the probability of sector attachment directly from the observed distribution of wages and worker attributes. This resolves the problem of attributing primary or secondary sector employment to everyone in a given industry or occupation. We then propose a direct test for involuntary confinement of workers to the secondary sector.

II. A Formal Test

How can we test the descriptive power of the dual market hypothesis without prior knowledge of the sector a person is in? Consider how we might proceed if people's earnings potential could be summarized by a single observable trait—for example, education—and an unobserved trait which was uncorrelated with education. In that case we could plot a scatter diagram of log wages and education. The standard view of the labor market holds that such a scatter diagram should resemble Figure 1. From dual market theory we would expect a scatter diagram similar to Figure 2. A straight-forward test of the theories would therefore entail plotting the scatter diagram and assessing whether it corresponds to either the human capital model or the dual market model.

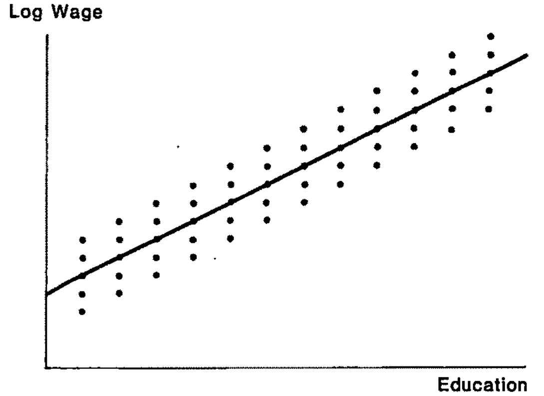


FIGURE 1. HYPOTHETICAL SCATTER PLOT—STANDARD HUMAN CAPITAL THEORY

Two problems complicate such an approach. First, wages are determined by many observable characteristics other than education. To control for all variables simultaneously, we would have to plot a scatter diagram for each subgroup in the sample. As the number of other variables increased, the number of observations on each diagram would decrease considerably. With a reasonable number of controls, the number of diagrams and the sparseness of observations would certainly make it impossible to discern any pattern. Second, even if we were able to plot all the scatter diagrams, we would still lack a formal mechanism for testing the hypotheses. Each researcher would be free to decide for him/herself whether the diagrams correspond more nearly to the predictions of human capital or dual market theory. These problems can be resolved by the use of the formal methods described in the following paragraphs.

The question of whether a plot looks more like Figure 1 or 2 can be rephrased: do two wage equations fit the data significantly better than one, and do the best-fitting equations fit the predictions of the dual market

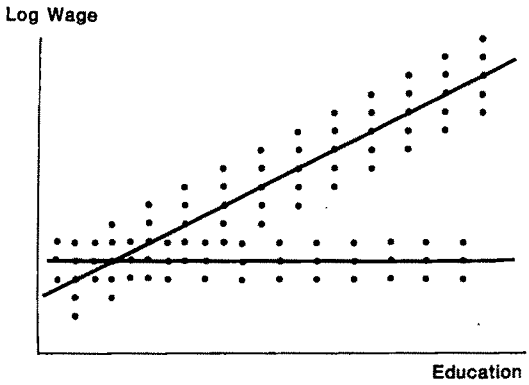


FIGURE 2. HYPOTHETICAL SCATTER PLOT—DUAL MARKET THEORY

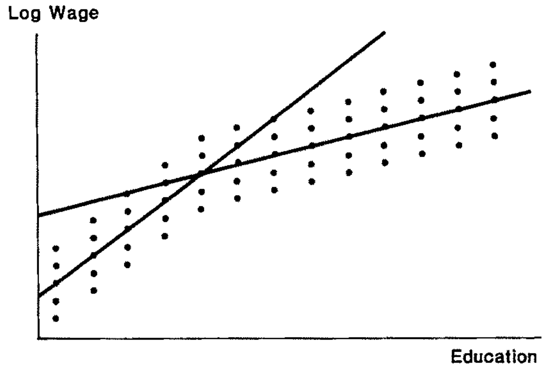


FIGURE 3. HYPOTHETICAL SCATTER PLOT—HUMAN CAPITAL THEORY WITH NONLINEAR WAGE-EDUCATION RELATION

hypothesis? We can imagine fitting first one, and then two lines by hand to Figure 2. To compare the explanatory power we might, for example, compute the distance from each point to the closest line. The reduction in the sum of squares going from one line to two would be much larger for Figure 2 than for Figure 1.

Of course, two equations having more explanatory power than one is not, by itself, a test of the dual market hypothesis. For example, two equations might have significantly more explanatory power than one for a scatter diagram such as Figure 3. However, there is no identifiable secondary market. Thus, in addition to requiring two equations to have significantly more explanatory power than one, we also require the best-fitting lines to have characteristics consistent with the dual market hypothesis. To correspond to the predictions of dual market theory, one wage equation should be upward sloping in schooling and experience, while the other equation should be flat with respect to human capital variables and below the other at most points. Since we are dealing with a sample of adult males, we also expect that there will be fewer observations associated with the low-wage line.

Formally, we may fit two wage equations using maximum likelihood techniques. Since we do not know a priori with which wage equation to compare an individual, we estimate a switching model with unknown regimes. To do this we must specify two wage equations and a third equation that predicts sector attachment, and estimate all three equations simultaneously. The likelihood function for this model can be found in the Appendix. Since the single-equation model is nested in the switching model, we may test the hypothesis that the two-equation model fits significantly better than the single-equation model by comparing the log-likelihood values for the two models. If we reject the single-equation hypothesis, we may examine the coefficients of the two wage equations to see if they fit the dual market hypothesis.

The existence of two sectors with different wage-setting mechanisms is fundamental to dual market theory, but it is not incompatible with human capital theory. While neoclassical economics tends to emphasize the development of models which are continuous and therefore tractable in calculus, if the technology were sharply discontinuous in the way suggested by Piore (1980b), no fundamental assumptions of mainstream economics would be violated. In this case, individuals would choose the sector of employment that maximized the expected present value of their lifetime utility.

The second tenet of dual market theory, that primary sector jobs are rationed, is less compatible with human capital theory. Dual market theory maintains that individuals cannot always choose the sector which they prefer—some workers who would prefer to be employed in the primary sector cannot find jobs there. As a general phenomenon

this would be incompatible with the standard neoclassical view. However, rationing as a general phenomenon is believed to be restricted to recession periods (Piore 1980a). During other periods only women and minorities are likely to experience rationing. This contention is no more troublesome than the widely acknowledged importance of race and sex discrimination in the determination of wages.

To test for the presence of noneconomic barriers to primary sector employment, we need to postulate a mechanism for allocating workers between the sectors in the absence of rationing. To begin, we assume that experience in one sector raises wages in the sector more than it raises wages in the other sector.[1] We also assume that workers will behave so as to maximize utility over their lifetime. Utility is assumed to be increasing with the net present value (NPV) of lifetime income. If we then assume that people's preferences with respect to the nonpecuniary aspects of jobs do not change over their lifetime, and that workers are perfectly informed about the characteristics of all jobs, we can conclude that workers will choose employment in one of the two sectors at the beginning of their careers and stay in that sector for their entire working life.[2]

If the nonpecuniary characteristics of the two sectors were similar, we would expect workers to pick the sector that yields the highest lifetime income. However, dual market theorists are unanimous in maintaining that the nonpecuniary aspects of secondary employment are inferior to those obtained in primary employment. On the other hand, starting wages in some secondary jobs may be higher than in the primary sector, and this could be attractive to a worker who plans to leave and enter the labor force frequently or change jobs often. In addition, secondary employers may be less concerned with lateness and absenteeism and the work pace may be slower in secondary jobs. Formally, we assume that workers will choose primary sector employment if the log of the NPV of their income stream in the primary employment exceeds the log of the NPV of secondary employment by more than an amount C—the additive inverse of the compensating differential for secondary employment. We may write the probability that a worker is employed in the primary sector (denoted P) as

$$(1)\quad P=\Pr\{\ln(NPV_p)-\ln(NPV_s)>C\}$$

where the subscripts p and s denote primary and secondary. To model the NPV in the two sectors, we write two wage equations:

$$(2)\qquad \ln(W_p)=XB_p+Ya_p+e_p;$$

$$(3)\qquad \ln(W_s)=XB_s+Ya_s+e_s,$$

where X is a vector of individual characteristics, Y is years of job experience, W_p is the wage received in the primary sector, e_p is a normally distributed error representing unobserved characteristics affecting the primary sector wage, and B_p and a_p are parameters. The terms W_s, e_s, B_s, and a_s are similarly defined for the secondary sector. Approximating the length of the individual's working life by infinity, and using (2) and (3), equation (1) becomes

$$(4)\quad P=\Pr\{X(B_p-B_s)+e_p-e_s+C'>0\},$$

where

$$(5)\quad C'=\ln((d-a_s)/(d-a_p))-C;$$

and d is the discount rate.

If we assume that C' is equal to a constant (C'') plus a normally distributed error term (i.e., people's preferences with respect to the nonpecuniary aspects of employment and their discount rates do not vary with observable characteristics (X)), we may test the hypothesis that people choose their sector of

[1]This assumption appears reasonable in light of recent empirical evidence on experience-earnings profiles. James Brown (1983, p. 20) shows that experience in other firms counts very little towards earnings for workers on their current jobs. The assumption entails the existence of sector-specific training. If some training is firm specific, it is ipso facto sector specific.

[2]It might be argued that young people, in particular, lack the necessary career information to make informed job choices. They may also have different preferences. Since we estimate our model on a sample of heads of households, these problems should not greatly affect our results.

employment to maximize their utility. We do this by estimating an equation to determine sector membership and testing the hypothesis that the coefficients on the X's are equal to $B_p - B_s$, or that the B_w's in

$$(6) \quad X(B_p - B_s + B_w) + C'' + e_p - e_s + e_w$$

are equal to zero.

It may not be reasonable to assume that preferences for the nonpecuniary aspects of primary or secondary employment are not related to any observed worker characteristics. If they are related, we would expect at least some of the B_w's to be different from zero even if workers are free to choose the sector they are employed in. In this case we may be able to find some X's that should not be related to tastes, or to suggest inequality constraints on the effects of certain characteristics on tastes. Specific tests of this type are proposed in Section III.

An intuitive explanation of this approach uses the example of race. Suppose that the lines fitting the scatter diagram in Figure 2 were the same for blacks and whites. Suppose further that the distribution of education was the same for the two groups, but that a higher proportion of blacks than of whites were scattered around the lower line. Under these circumstances, we would conclude that either blacks are less averse to secondary employment than are whites, or that blacks face discrimination in obtaining primary jobs. Supplementary evidence would support the latter explanation.

The data used in this study are drawn from the thirteenth wave (1980) of the *Panel Study of Income Dynamics*. We limited the sample to men who were heads of households, working more than 1000 hours in the previous year, did not work in government and for whom data on education and marital status were available. Estimates were obtained for both the full sample (2812 cases) and with only members of the Survey Research Center sample (1696 cases).

III. Results

Table 1 presents the results for *OLS* estimation and the dual market model. Since the results for the samples are similar, we discuss only the restricted sample here. The *OLS* results are similar to those obtained by other researchers. The return to schooling is about 6 percent while the return to experience is about 1 percent. Whites receive wages about 13 percent higher than nonwhites, holding other factors constant. Workers living in an SMSA earn wages almost 20 percent higher than equivalent workers outside an SMSA, and workers who have never been married earn considerably less than other workers. All the coefficients are significant at conventional levels.

The second part of Table 1 tells a very different story. The primary sector wage equation resembles the *OLS* equation, but there are some striking differences. Most notably, the white-nonwhite differential falls to zero (although it is measured very imprecisely). In addition, the effect of living in an SMSA declines and the returns to schooling and experience increase somewhat.

The secondary sector wage equation contrasts sharply with the *OLS* equation. None of the coefficients is statistically significant at conventional levels. We cannot reject the hypothesis that the secondary sector wage equation is completely flat. The return to experience (which is measured quite precisely) is essentially zero. Further, the secondary sector wage equation is almost everywhere below the primary sector. For a nonwhite living in an SMSA who has never been married and has a sixth-grade education, the predicted primary sector wage is greater than the secondary wage after one year's experience. For all other workers, except those with less education, the predicted primary sector wage is always higher than the predicted secondary sector wage. Since the coefficients of the secondary sector wage equation are measured imprecisely, it might be presumed that, in fact, there is only one labor market. However, using a likelihood ratio test, we can easily reject the single labor market (*OLS*) model at any conventional level of significance.[3] Two wage equations fit the data considerably better than one.

[3]Twice the difference between the log-likelihood values for the two models is 177. Although the single equation model is nested in the switching model, when

TABLE 1[a]

Variable	Mean	OLS	Switching Model Primary	Switching Model Secondary	Switching Model Switch
Restricted Sample					
Constant	1.00	.874	.996	1.32	−.006
		(.075)	(.297)	(3.33)	(.574)
SMSA	0.67	.197	.112	.197	.361
		(.025)	(.060)	(1.28)	(.158)
Never Married	0.08	−.305	−.261	−.244	−.157
		(.044)	(.055)	(.580)	(.354)
School	12.7	.059	.067	−.003	.020
		(.005)	(.005)	(.072)	(.031)
White	.91	.134	.008	−.192	.796
		(.040)	(.166)	(2.73)	(.328)
Experience	18.4	.010	.013	.001	
		(.001)	(.001)	(.002)	
Covariance with Switching Error			.068	−.009	
			(.389)	(4.42)	
Standard Error		.477	.374	.381	b
Log-Likelihood		−1151.4		−1062.9	
Full Sample					
Constant	1.00	.760	.982	1.27	−.389
		(.051)	(.108)	(.636)	(.379)
SMSA	0.69	.194	.078	.073	.526
		(.020)	(.036)	(.452)	(.144)
Never Married	0.10	−.265	−.286	−.268	.238
		(.031)	(.047)	(.263)	(.338)
School	12.07	.063	.069	.006	.037
		(.004)	(.005)	(.034)	(.024)
White	0.67	.180	.006	−.139	.885
		(.020)	(.059)	(.781)	(.190)
Experience	17.9	.010	.014	.000	
		(.001)	(.001)	(.002)	
Covariance with Switching Error		–	.155	−.019	b
			(.084)	(1.18)	
Standard Error		.471	.392	.373	
Log-Likelihood		−1875.3		−1772.9	

[a]Standard errors are shown in parentheses; dependent variable is log hourly wage.
[b]Normalized to 1.

the switching equation model is constrained to yield the single-equation model, several parameters are unidentified. This problem complicates the calculation of the degrees of freedom. In addition, it is possible that the asymptotic likelihood ratio statistic does not have a *chi*-squared distribution. However, Monte Carlo tests (Steven Goldfeld and Richard Quandt, 1976) suggest that setting the degrees of freedom equal to the number of constraints plus the number of unidentified parameters yields a conservative test using the *chi*-squared distribution. For our problem, this computation yields fourteen degrees of freedom. The 1 percent critical value for the *chi*-squared distribution with fourteen degrees of freedom is 29.14—far smaller than our computed likelihood ratio test statistic.

Thus we can reject the single labor market model and cannot reject the predictions of dual market theory that there are no returns to education or experience in the secondary sector. As noted above, this characterization of the market, while not commonly assumed in mainstream economics, is not incompatible with it. A more crucial aspect of dual market theory is the assumption that primary sector jobs are rationed. Testing this assumption entails testing constraints on the switching equation. Using the restricted sample, we were unable to get the constrained likelihood

function to converge. Since a Wald test of the constraints is not invariant with respect to the choice of normalization,[4] a likelihood ratio test is preferable. In the following paragraphs, we report the results of likelihood ratio tests performed on the full sample.

If workers were free to choose between the sectors and tastes for the nonpecuniary aspects of employment were not related to the location of a worker's residence, his marital status, education, or race, we would expect the coefficients of these variables in the switching equation to equal the difference between the coefficients in the two wage equations. However, it is probably not reasonable to expect workers' preferences with respect to nonpecuniary job attributes to be independent of these variables. For example, we would not be surprised to find that workers outside of SMSAs required less of a compensating differential to get them to take secondary work since they may often be engaged in agricultural labor. We therefore test the hypothesis that B_w in equation (6) equals zero for school, white, and never married. Twice the difference between the log-likelihoods for the constrained and unconstrained models is 14.92. The 1 percent critical value for the *chi*-square with two degrees of freedom is 9.21.[5] The hypothesis of free choice is easily rejected.

Since married workers may have a greater desire for stable primary work, we may want to restrict our attention to the coefficients of education and race in the switching equation. We can reject the hypothesis that these coefficients are both equal to the difference between their corresponding coefficients in the primary and secondary wage equations ($\chi^2 = 14.56$, 1 percent critical value for one degree of freedom = 6.63). We are left with three potential explanations for our results. First, highly educated workers prefer secondary employment more than less-educated workers. This hypothesis seems unlikely. We would expect more-educated workers to be more averse to the poor working conditions of secondary employment. Shulamit Kahn (1983) finds that the demand for occupational safety increases with education. A second explanation is that blacks are less averse to secondary jobs than are whites, but this runs counter to evidence that blacks are more likely to support unions in representation elections (Henry Farber and Daniel Saks, 1980; Dickens, 1983), are less likely to quit a job (W. Kip Viscusi, 1979), and have greater demand for occupational safety than equivalent whites (Kahn). Primary jobs are more likely to be unionized, and offer more stable employment and better job safety. If we cannot accept these other two explanations, we are forced to conclude that blacks face noneconomic barriers to employment in the primary sector. At the present time there is no formal way of establishing which of these three explanations is correct. However, since the first two hypotheses appear to be inconsistent with other studies of the demand for job quality, the most reasonable explanation is the last—blacks are discriminated against when seeking primary employment.

If we accept the dual market hypothesis, we may use the model to determine the composition of the primary and secondary sectors.[6] According to our estimates, about 12 percent of working male heads of households are employed in the secondary sector.

[4] In the unrestricted model it is not possible to simultaneously identify all the coefficients of the switching equation and its error variance. This is a problem common to all discrete dependent variable estimation. Thus any one restriction on the coefficients of the switching equation cannot be tested as it would only constitute a normalization. It is possible to perform a Wald test if there is more than one constraint, but the test is not invariant to the normalization chosen. In all cases reported below, the results of the Wald test were inconclusive since the Wald test rejected the null hypothesis for some reasonable normalizations but not for others.

[5] We are imposing three constraints, but we also relax the normalization that the variance of the switching equation equals one. Thus there are only two degrees of freedom.

[6] A straightforward application of the Bayes theorem gives the result that the probability that worker i is in the primary sector conditional on the observed wage and personal characteristics is

$$\frac{\Pr(\varepsilon_{wi} > -Z_i\Gamma | Z_i, X_i, \varepsilon_{pi}) f(\varepsilon_{pi})}{\Pr(\varepsilon_{wi} > -Z_i\Gamma | Z_i, X_i, \varepsilon_{pi}) f(\varepsilon_{pi}) + \Pr(\varepsilon_{wi} \leq -Z_i\Gamma | Z_i, X_i, \varepsilon_{si}) f(\varepsilon_{si})}$$

where the notation is described in the Appendix. The percent of workers is estimated by computing the average value of the probability of primary sector attachment for all workers.

TABLE 2—COMPOSITION OF SAMPLE AND SECONDARY SECTOR

		Unrestricted Model[a]		Restricted Model[a]	
	Percent of Sample in Category	Percent of Secondary Sector Workers in Category	Percent of Workers in each Category in Secondary Sector	Percent of Secondary	Percent in Secondary
SMSA	66.9	55.3	10.3	46.7	7.9
Not SMSA	33.1	44.7	16.8	53.3	18.3
Married	91.8	89.5	12.1	90.9	11.2
Not Married	8.2	10.5	15.9	9.1	12.6
Education < 12	19.9	28.4	17.8	26.1	14.9
Education = 12	39.8	40.2	12.5	39.9	11.4
Education > 12	40.3	31.4	9.7	34.0	9.6
White	90.6	76.3	10.5	84.9	10.6
Nonwhite	9.4	23.7	31.1	15.1	18.2
Age					
< 25	13.9	21.4	19.1	20.6	16.8
25–29	20.8	19.2	11.5	19.2	10.5
30–39	29.8	22.6	9.4	23.2	8.8
40–49	14.4	12.6	10.8	12.5	9.8
50–59	15.7	11.6	9.2	11.3	8.2
60+	5.3	12.6	29.5	13.2	28.1
Total			12.4		11.3

[a]Numbers in columns for both models were estimated using the formula described in fn. 6.

This seems large, especially since we would expect a sample containing teenagers, women, and the unemployed to have a higher proportion of secondary workers. Table 2 shows the makeup of the sample and the secondary market. It also shows the percent of each type of worker in the secondary market. Since many of the parameters of the switching model are estimated with a great deal of error, we also estimated a restricted model (parameter estimates are shown in Table 3) where the wage equation in the secondary sector was constrained to be flat and education and marital status were removed from the switching equation. (A likelihood ratio test fails to reject the constraints at the .1 level.)[7] Both models show the same pattern evident in the parameters of the switching equation: workers in SMSAs, married workers, more educated workers, and whites are less likely to be in the secondary sector. In addition, heads of household less than 25-years old or 60-years or older are disproportionately in the secondary sector.

Finally, we examine how sharply the model distinguishes between workers in the primary and secondary sectors. Figure 4 shows the distribution of predicted probabilities of being in the primary market. The distribu-

TABLE 3—ESTIMATES FOR RESTRICTED MODEL[a]

Variable	Primary	Secondary	Switching
Constant	.887	1.22	.503
	(.073)	(.093)	(.281)
SMSA	.108	[c]	−.537
	(.026)		(.132)
Never married	−.288	[c]	[c]
	(.037)		
School	.069	[c]	[c]
	(.004)		
White	.083	[c]	.433
	(.043)		(.227)
Experience	.013	[c]	[c]
	(.001)		
Log-Likelihood	−1069.1		
SE	.3773	.4098	[b]

[a]Standard errors are shown in parentheses; dependent variable: log of hourly wage.
[b]Normalized to 1.
[c]Constrained to zero.

[7]Again we note that the measured returns to schooling and education in the primary sector are larger than in the *OLS* equation. Also, the "discrimination coefficient" is roughly 40 percent smaller.

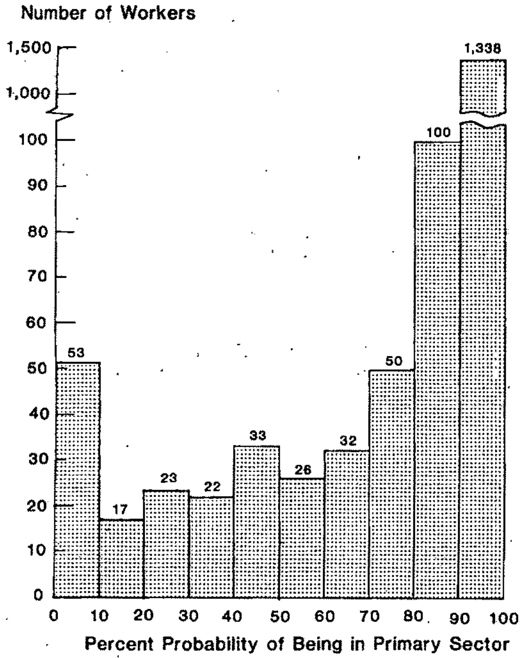


FIGURE 4. PREDICTED PROBABILITY OF BEING IN PRIMARY SECTOR—FREQUENCY DISTRIBUTION OF NUMBER OF WORKERS (SAMPLE SIZE: 1,696)

tion is distinctly bimodal, with the two modes of 0–10 and 90–100 percent probability. There is a large group of workers who are clearly identifiable as being in the secondary sector, and a larger group with a high probability of being primary workers. It appears that there is a distinct secondary sector which the model can identify.

IV. Conclusions

Our results provide strong support for two of the basic tenets of dual market theory: there are two distinct sectors of the labor market with different wage-setting mechanisms, and there is a queue for primary sector jobs. We believe that our approach and results represent a considerable advance over previous research in this area. By allowing the distribution of wages and worker attributes to determine our "assignment" of workers to sectors, we avoid the problems of arbitrariness and sample selection bias that complicate the interpretation of earlier research. In addition, our approach allows us to estimate the size and composition of the secondary work force in a noncircular manner.

Of course, we cannot exclude other interpretations of these results that postulate different distributions of the error term, or some unusual nonlinear functional form for the wage equation. While we cannot deny these possibilities, we suggest that in the absence of our results, such a distribution would not be suggested. It was dual market theory that led to our test, and the results therefore tend to corroborate that theory.

Given the strength of the reactions (deifying or executing the messenger) of some of the individuals with whom we discussed preliminary results, it is important to take stock of exactly what it is that we have and have not shown. Piore (1983) suggests that the strength of opposition to dual market theory is due, in part, to the use of participant observer techniques rather than econometric techniques that are more common in mainstream economics. We have shown that the dual market hypothesis can be derived and supported from standard data and statistical techniques. It is, however, unlikely that standard approaches would have uncovered labor market duality, a fact that suggests that there is a role for other methods in mainstream economics.

On the other hand, the fact that we can test dual market theory using mainstream techniques suggests that the two theories are not as incompatible as would appear from the antagonisms in the profession. We have already suggested that neoclassical economics makes few assumptions regarding the nature of technology. It is relatively straightforward to develop a model in which a high fixed cost/low variable cost technology is used in the "stable" demand sector and a low fixed cost/high variable cost technology is used to accommodate fluctuations in demand. Piore (1980b) gives a verbal description of such a theory and Elie Appelbaum and Chin Lin (1982) present a formalization. It is a direct consequence of human capital theory that workers and firms will invest little in firm-specific training if the worker is

not expected to remain with the firm for very long. Thus the existence of two markets with distinct wage profiles can be easily accommodated by mainstream theory.

Similarly, while when first proposed, the view that there is a queue for primary sector jobs may have appeared to be incompatible with neoclassical theory, there are an increasing number of imperfect information models which imply that there can be a queue for jobs. In particular, Andrew Weiss (1980), Steve Stoft (1982), Carl Shapiro and Joseph Stiglitz (1984), and Samuel Bowles (1985) have developed models in which job queues arise in firms in which there are unobserved skills or effort. Thus there could well be a queue for primary sector jobs. If there are few skill differences in secondary sector jobs, there would be no queues for them.

While these models are compatible with queues, we have presented evidence that rather than allocating jobs randomly, primary sector employers discriminate against nonwhites. This may appear to be incompatible with neoclassical economics. However, discrimination is an anomaly which remains to be explained, whether or not one accepts dual market theory. In fact, these results may help to explain the existence and persistence of discrimination. According to the point estimates presented in the last section, more than 40 percent of white-nonwhite wage differences can be explained in the restricted model by the fact that nonwhites are crowded into the secondary sector, while in the unrestricted model the within-sector differential is zero. If a queue for primary jobs exists because wages are at least partially socially determined (George Akerlof, 1982), primary employers with a "taste" for discrimination may indulge it by hiring fewer nonwhites from the queue without sacrificing profits. No economic incentive exists for the elimination of this sort of discrimination. If, on the other hand, the queue results from "economic" causes such as unobservable skills or effort, the usual result that competition should eliminate discrimination applies, and its existence continues to present theoretical difficulties.

We do not wish to imply that there are no incompatibilities between dual market theory and neoclassical economics. For example, dual market theorists have generally assumed that preferences are endogenous, a position strongly resisted by most mainstream economists despite some exceptions.

Perhaps more important, dual market theorists have developed very elaborate theories of the origin and operation of labor market institutions which are rich in historical detail.[8] These descriptions are quite remote in many ways from the neoclassical description of the labor market. However, we have not attempted to test these aspects of dual market theory (Reich, 1984, does).

Finally, we call the reader's attention to the title of this paper. We have chosen to refer to our work as a test of dual market theory rather than as a test of human capital theory because, in our view, dual market theory is not necessarily incompatible with standard neoclassical analysis. Our results therefore point to the need for additional work to understand the origins of these institutions rather than to abandon the neoclassical model of the labor market. In addition, our results point to the value of noneconometric techniques for uncovering and understanding labor market institutions.

Appendix

Estimation of the Switching Model with Unknown Regimes

Consider the system composed of wage equations for each sector and an equation determining "tendency to be in the primary sector":

$$\ln W_i = X_i \beta_p + \varepsilon_{pi}, \tag{A1}$$

$$\ln W_i = X_i \beta_s + \varepsilon_{si}, \tag{A2}$$

$$y_i^* = Z_i \Gamma + \varepsilon_{wi}, \tag{A3}$$

where W_i is the individual's wages, X_i and Z_i are vectors of explanatory variables, β_p, β_s, and Γ are vectors of parameters, ε_p, ε_s, and ε_w are normally distributed error terms,

[8] For example see Piore (1980a), Richard Edwards (1979), and Gordon, Edwards, and Michael Reich (1982).

and y^* is a latent variable measuring tendency to be in the primary sector. Equation (A1) is the wage equation if the individual is in the primary sector; (A2) is the wage equation if the individual is in the secondary sector; and (A3) is the switching equation.

We do not observe y^*. However, if $y^* > 0$, the individual's wage is determined by (A1); otherwise it is determined by (A2). Equivalently, the individual works in the primary sector if and only if

$$\text{(A4)} \qquad \varepsilon_{wi} > -Z_i\Gamma.$$

The likelihood function for the problem is therefore given by

$$\text{(A5)} \quad \Pr(\varepsilon_{wi} > -Z_i\Gamma | Z_i, X_i, \varepsilon_{pi})\cdot f(\varepsilon_{pi}) + \Pr(\varepsilon_{wi} \leq -Z_i\Gamma | Z_i, X_i, \varepsilon_{si})\cdot f(\varepsilon_{si})$$

where $f(\,)$ is the density of the error ε_p or ε_s. If we assume that ε_p, ε_s, and ε_w are normally distributed, the log-likelihood is thus:

(A6)

$$\sum_{i=1}^{N} \ln\left\{\left[1-\Phi\left(\frac{-Z\Gamma-\frac{\sigma_{pw}}{\sigma_{pp}}\varepsilon_{pi}}{\left(1-\frac{\sigma_{pw}^2}{\sigma_{pp}}\right)^{.5}}\right)\right]\cdot\phi(\varepsilon_{pi},\sigma_{pp}) + \Phi\left(\frac{-Z_i\Gamma-\frac{\sigma_{sw}}{\sigma_{ss}}\varepsilon_{si}}{\left(1-\frac{\sigma_{sw}^2}{\sigma_{ss}}\right)^{.5}}\right)\cdot\phi(\varepsilon_{si},\sigma_{ss})\right\}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the normal density and cumulative distribution, respectively, and σ_{jk} is the covariance of ε_{ji}; and ε_{ki}; σ_{ww} is normalized to equal one. Maximum likelihood estimates for Γ, β_p, β_s, and the σ's can be obtained using standard search algorithms provided that care is taken to prevent the program from iterating into regions for which the likelihood function is unbounded.

It is easy to see that if β_p equals β_s, and σ_{pw} equals σ_{sw}, then ε_{pi} equals ε_{si} and the likelihood function reduces to the standard normal density. It is therefore possible to test for the existence of two regimes by comparing the log-likelihood values for *OLS* and unknown regime estimates by performing a likelihood ratio test.

The likelihood functions used here were maximized using the Ernst Berndt et al. (1974) algorithm. While the nonlinearity of the system made convergence difficult, we did not experience any difficulties with unboundedness. All unconstrained specifications converged to interior solutions from *OLS* starting values.

REFERENCES

Akerlof, George A., "Labor Contract as Partial Gift Exchange," *Quarterly Journal of Economics*, November 1982, *97*, 543–69.

Appelbaum, Elie and Lin, Chin, "Ex Ante and Ex Post firms and Stochastic Equilibrium Price Fluctuations," mimeo., University of California-Irvine, 1982.

Berger, Suzanne and Piore, Michael J., *Dualism and Discontinuity in Industrial Societies*, New York: University of Cambridge Press, 1980.

Berndt, E. K. et al., "Estimation and Inference in Non-Linear Structural Models," *Annals of Economic and Social Measurement*, October 1974, *3*, 653–65.

Bibb, Robert and Form, William H., "The Effects of Industrial, Occupational and Sex Stratification on Wages in Blue-Collar Markets," *Social Forces*, June 1977, *55*, 974–96.

Bowles, Samuel, "The Production Process in a Competitive Economy: Walrasian, Neo-Hobbesian, and Marxian Models," *American Economic Review*, March 1985, *75*, 16–36.

Brown, James N., "Are Those Paid More Really No More Productive? Measuring the Relative Importance of Tenure Versus On-The-Job Training in Explaining Wage Growth," mimeo., Princeton University, July 1983.

Buchele, Robert, (1976a) "Jobs and Workers: A Labor Market Segmentation Perspective on the Work Experience of Young Men," unpublished doctoral dissertation, Harvard

University, May 1976.

______, (1976b) "Jobs and Workers: A Labor Market Segmentation Perspective on the Work Experience of Middle-Aged Men," unpublished paper submitted to the Secretary of Labor's Conference on the National Longitudinal Survey of the Pre-Retirement Years, Boston, 1976.

Cain, Glenn, "The Challenge of Segmented Labor Market Theories to Orthodox Theory," *Journal of Economic Literature*, December 1976, *14*, 1215–57.

Carnoy, M. and Rumberger, R., "Segmentation in the U.S. Labor Market: Its Effects on the Mobility and Earnings of Whites and Blacks," *Cambridge Journal of Economics*, June 1980, *4*, 117–32.

Dickens, William T., "The Effect of Company Campaigns on Certification Elections: Law and Reality Once Again," *Industrial and Labor Relations Review*, July 1983, *36*, 560–75.

Doeringer, Peter B. and Piore, Michael J., *Internal Labor Markets and Manpower Analysis*. Lexington: Lexington Books, 1971.

Edwards, Richard, *Contested Terrain*, New York: Basic Books, 1979.

Farber, Henry S. and Saks, Daniel H., "Why Workers Want Unions: The Role of Relative Wages and Job Characteristics," *Journal of Political Economy*, April 1980, *88*, 346–69.

Goldfeld, Steven M. and Quandt, Richard E., "Techniques for Estimating Switching Regressions," in their *Studies in Non-Linear Estimation*, Cambridge: Ballinger, 1976.

Gordon, David M., "Class, Productivity and the Ghetto," unpublished doctoral dissertation, Harvard University, 1971.

______, Edwards, Richard C., Reich, Michael S., *Segmented Work, Divided Workers: The Historical Transformation of Labor in the United States*. Cambridge: Cambridge University Press, 1982.

Hodson, Randy D., "Labor Force Participation and Earnings in the Core, Peripheral, and State Sectors of Production," M.A. thesis, Department of Sociology, University of Wisconsin-Madison, 1977.

Kahn, Shulamit, "Occupational Safety and Worker Preferences," unpublished doctoral dissertation, MIT, 1983.

Leigh, Duane E., "Occupational Advancement in the Late 1960s: An Indirect Test of the Dual Labor Market Hypothesis," *Journal of Human Resources*, Spring 1976, *11*, 155–71.

Oster, Gerry, "A Factor Analytic Test of the Theory of the Dual Economy," *Review of Economics and Statistics*, February 1979, *61*, 33–51.

Osterman, Paul, "An Empirical Study of Labor Market Segmentation," *Industrial and Labor Relations Review*, July 1975, *28*, 508–23.

Piore, Michael J., (1980a) "Dualism as a Response to Flux and Uncertainty," in Suzanne Berger and his *Dualism and Discontinuity in Industrial Societies*, New York: Cambridge University Press, 1980.

______, (1980b) "The Technological Foundations of Dualism," in Suzanne Berger and his *Dualism and Discontinuity in Industrial Societies*, New York: Cambridge University Press, 1980.

______, "Labor Market Segmentation: To What Paradigm Does It Belong?," *American Economic Review Proceedings*, May 1983, *73*, 249–53.

Reich, Michael, "Segmented Labour: Time Series Hypotheses and Evidence," *Cambridge Journal of Economics*, March 1984, *8*, 63–81.

Rosenberg, Samuel, "An Empirical Test of the Dual Labor Market Hypothesis," unpublished manuscript, Williams College, 1976.

______, "A Survey of Empirical Work on Labor Market Segmentation," Working Paper 138, Department of Economics, University of California-Davis, November 1979.

Schiller, Bradley R., "Relative Earnings Mobility in the United States," *American Economic Review*, December 1977, *67*, 926–41.

Shapiro, Carl and Stiglitz, Joseph E., "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, *74*, 433–44.

Stoft, Steve, "Cheat Threat Theory: An Explanation of Involuntary Unemployment," mimeo., Boston University, May 1982.

Viscusi, W. Kip, *Employment Hazards: An*

Investigation of Market Performance, Cambridge: Harvard University Press, 1979.

Weiss, Andrew, "Job Queues and Layoffs in Labor Markets with Flexible Wage Expectations," *Journal of Political Economy*, June 1980, *88*, 526–38.

Wright, Eric O., *Class Structure and Income Determination*, New York: Academic Press, 1979.

Zucker, Lynne G. and Rosenstein, Carolyn, "Taxonomies of Institutional Structure: Dual Economy Reconsidered," *American Sociological Review*, December 1981, *46*, 869–83.

Unbalanced Growth Revisited: Asymptotic Stagnancy and New Evidence

By WILLIAM J. BAUMOL, SUE ANNE BATEY BLACKMAN, AND EDWARD N. WOLFF*

Some years ago, Baumol (1967) presented a model of unbalanced growth in which an oversimplified economy was divided into productivity growth sectors, one "stagnant" and one "progressive." It was argued that relative costs and prices in the stagnant sector would tend to rise persistently and cumulatively, and that *if* the output proportions of the two sectors happened to remain fairly constant, the share of the economy's inputs used by the stagnant sector and the share of consumer expenditure devoted to outputs of the stagnant sector must both rise toward 100 percent. Finally, it was concluded that the net result must be a *ceteris paribus* decline in the economy's overall productivity growth rate.

Since then a variety of pertinent empirical materials and some further analysis have suggested that the model needs modifications, some of them of interest in themselves. But the behavior of prices, input-use patterns, and consumer outlays have followed the model's scenario to a remarkable degree.

I. Manifest Destiny of Relative Costs and Sectoral Inputs

In this paper we show that Baumol's earlier equation of the service sector of reality with the stagnant sector of the model requires modification. But there *is* a subclass of the services that is a better approximation to the model's stagnant activities. We also introduce a third set of economic activities, that we label "asymptotically stagnant," which are neither completely stagnant nor progressive. They use, in fairly fixed proportions, some inputs from the progressive sector and some from the stagnant sector. We will show that in their initial phases such activities are often outstanding in their rapid productivity growth and declining costs. However, with the passage of time, the cost and price behavior of these asymptotically stagnant activities *necessarily* approaches that of the stagnant sector.

We will also examine the empirical evidence relating to the model, showing that:

(*i*) In real terms, there happens to have been little shift in output shares between manufacturing and the services, not only with time, but with increasing wealth as one goes from less developed to industrialized countries. The model does not predict this, but the trend is not inconsistent with it.

(*ii*) As the model predicts, with these constant output proportions there was a marked simultaneous rise in relative prices and share of total expenditure on the services both with the passage of time and with increased industrialization.

(*iii*) The service sector happens to contain some of the economy's most progressive activities as well as its most stagnant.

(*iv*) As the model predicts, the U.S. labor force has been absorbed predominantly by the stagnant subsector of the services rather than the services as a whole.

(*v*) Television broadcasting and electronic data processing are examples of asymptotically stagnant activities, and the empirical budget and cost patterns for these activities are perfectly consistent with the model's predictions.

II. Basic Results on Stagnant and Progressive Outputs

Before summarizing the basic propositions to be evaluated empirically, we emphasize two crucial qualifications. First, the model is

*Baumol: Princeton and New York universities; Blackman: Princeton University; Wolff: New York University. We are extremely grateful to the Division of Information Science and Technology of the National Science Foundation, the Exxon Education Foundation, the Fishman-Davidson Center for the Study of the Service Sector, and the C. V. Starr Center for Applied Economics at New York University for support of the research reported here. We also thank David Dollar for his valuable suggestions.

obviously a gross oversimplification. Outputs, firms, and industries do *not* fall into black and white categories of stagnancy and progressivity—they are all shades of gray. Even the most stagnant sectors of the economy have undergone some technological change, varying from one period to another. Second, an activity which is, say, relatively stagnant need not stay so forever. It may be replaced by a more progressive substitute, or it may undergo an outburst of innovation previously thought very unlikely. Thus, there may be radical changes in the time paths predicted by the model. History shows the folly of predicting that some field of endeavor is beyond human inventiveness. When we speak of manifest destiny here, our claim is more modest. We merely maintain that things must go as predicted only so long as there is no major qualitative change in the distribution of innovation among industries.

The earlier paper on unbalanced growth provided some basic propositions whose proofs can now be generalized considerably. However, here we merely restate them and a few corollaries:

1) With the passage of time, the cost per unit of a consistently stagnant product (for example, live concerts) will rise monotonically and without limit relative to the cost of a consistently progressive product (for example, watches and clocks).

The reason for this phenomenon, which has been called the cost disease of the stagnant services, is obvious—the growing relative productivity of a more progressive output means that it will use relatively smaller and smaller input quantities per unit of output as time passes.

2) If the output ratio of a stagnant to a progressive product (the number of concerts performed divided by the number of watches produced) happens to remain constant or does not fall, the share of the combined inputs used by the stagnant activity must rise without limit.

This, too, is a tautology, since the progressive output must by definition employ relatively less and less input per unit of output, and the relative decline in its input use must compound with the passage of time.

3) If relative prices correspond to relative unit costs and if the ratio of the stagnant to the progressive output does not fall, then relative expenditure on the stagnant product must rise monotonically with time.

An example will make this clear and suggest the magnitudes that may be involved. Between the 1670's and the 1970's, the output per watchmaker in Geneva is estimated to have risen from about 12 watches to over 1,200 watches per year. Purcell wrote *Dido and Aneas* in the 1680's and today it takes as many person-hours and instruments to perform *live* as it did then. Hence, if the ratio of watches produced to performances of Dido had remained exactly the same, both the relative input quantities devoted to the musical performance and the relative expenditures on the performances must have risen about one-hundredfold.

From all this we conclude:

4) In an economy in which the productivity growth rates of the different sectors are unequal, it is impossible for both the output ratios and the input ratios to remain constant.

III. On Asymptotically Stagnant Activities

We come now to our third type of activity which was not included in the earlier 1967 model. These are the asymptotically stagnant activities like TV broadcasting and data processing that we think of as outstandingly progressive, but whose progressivity, as we will show, carries the seeds of its own destruction.

A pure asymptotically stagnant activity is one that uses in fixed proportions one group of inputs produced by progressive activities and another set of inputs produced by stagnant activities. A prime example is television broadcasting with, roughly, one hour of its progressive component (electronic transmission) required for one hour of its stagnant input (performance or program production). Characteristically, these are "high tech" industries, at the frontier of technical progress.

These activities are noteworthy for their behavior patterns. In their early stages, when progressive inputs dominate their budgets, their costs and prices fall rapidly, like those of progressive activities. Later, their fixed input proportions and the rapid fall in the relative prices of their progressive inputs *in-*

evitably give the stagnant inputs an ever-rising share of the total budget of the asymptotically stagnant activity, as a simple matter of arithmetic. For example, if the progressive input's cost is initially 80 percent of the budget and falls 25 percent per year, while the stagnant input is 20 percent of the budget and rises 6 percent per year (these, as we will see, are approximate figures for data processing), a pocket calculator will confirm that in just about ten years the budget proportions *must* be reversed, with the stagnant output now about 80 percent of the total. Third, as the stagnant component *must* come to dominate the activity's budget, its output cost and price must approach those of its stagnant component, and therefore have to rise, succumbing to the cost disease. Finally, the date when the activity sheds its progressive characteristics comes more quickly the more rapid the decline in the price of its progressive component. This is so because the more spectacularly successful is productivity enhancement in the production of the progressive inputs, the more rapidly they will distinguish themselves as significant components of the asymptotically stagnant activity's budget and, consequently, the more rapidly the relative cost of this activity must begin to rise.

These results can also all be derived via formal mathematics, but this is not the place to do so.

IV. Empirical Evidence from the U.S. Economy

We turn now to our empirical evidence—to test the implications of the basic model of unbalanced growth, and the asymptotic stagnancy construct. The first of these tasks requires classification of the actual sectors of the economy into progressive and stagnant categories, a division that is inevitably somewhat arbitrary. We base the classification on input and output data for the U.S. economy for 1947–76, since consistent national account data and input-output tables are available. A variety of measures of productivity growth rates were used to test the sensitivity of our classification scheme.

In Table 1, column 1 shows calculations of annual (compounded) rates of labor productivity growth using the official *National Income and Product Accounts*... (BEA, 1981).[1] The corresponding sectoral productivity concept is gross product originating (*GPO*) per person employed, and that of aggregate productivity is the ratio of gross domestic product (*GDP*) to total persons employed. The average annual rate of aggregate productivity growth was 2.16 percent over the period. Sectoral rates of productivity growth ranged from a high of 5.42 percent in communications and broadcasting, a service sector, to a low of −0.51 percent in government enterprises. Though there is a fairly wide spread in sectoral rates of productivity growth, there also appears to be a sharp break between the construction sector at 1.66 percent and the narrowly defined "general services" sector at 0.93 percent. By this criterion and these data, four sectors are stagnant: services (0.93 percent); finance and insurance (0.50 percent); government industry (0.31 percent); and government enterprises (−0.51 percent). Productivity growth in the remaining sectors was fairly rapid, putting them in the progressive group. Note that this group includes three service sectors: communications; trade; and real estate.[2]

The second column of Table 1 uses gross domestic output (*GDO*) in constant dollars as its sectoral output and number of persons employed as its labor input. *GDO* in constant dollars, an input-output concept, equals gross value of a sector's output or sales deflated by the *sectoral* price deflator. The new

[1] Here, as the total value of goods and services produced domestically, irrespective of ownership, *GDP* is actually preferable. The level of industry disaggregation was determined by the available statistics for the period. The output variable is gross product originating (*GPO*) in constant (1972) dollars. *GPO* in constant dollars is defined as the difference between the deflated value of output and the deflated value of interindustry inputs. The input concept is "persons engaged in employment" (L), defined as the sum of the number of full-time-equivalent employees and self-employed workers. This is perhaps the best available measure of labor input.

[2] The real estate data must be interpreted cautiously, since part of the "output" is the rent imputed to owner-occupied housing. However, where imputed rent enters official *GNP* and *GDP* statistics, the reported rate of productivity growth in real estate is the appropriate datum.

TABLE 1—AVERAGE ANNUAL RATE OF PRODUCTIVITY GROWTH BY SECTOR, 1947–76[a]

	Measure			
Industry	GPO/L (1)	GDO/L (2)	ρ (3)	λ (4)
1. Agriculture	3.59	4.47	1.56	3.95
2. Mining	2.70	2.76	0.08	1.38
3. Construction	1.66	1.19	−0.34	1.49
4. Manufacturing-Durables	2.52	2.80	0.58	3.08
5. Manufacturing-Nondurables	3.21	3.23	0.41	2.56
6. Transportation and Warehousing	1.74	2.74	0.68	2.42
7. Communication and Broadcasting	5.42	5.50	3.99	5.21
8. Utilities	4.96	4.77	1.53	2.96
9. Trade		2.17	1.09	2.19
a. Wholesale Trade	2.37			
b. Retail Trade	1.99			
10. Finance and Insurance	0.50	0.31	−0.27	0.57
11. Real Estate	2.72	3.10	1.21	4.86
12. General Services	0.93			
a. Hotels, Personal and Repair (except auto)		1.37	−0.31	1.35
b. Business and Professional Services		1.70	0.83	2.30
c. Auto Repair and Services		1.45	−0.84	1.04
d. Movies and Amusements		0.99	−0.56	0.64
e. Medical, Educational and Nonprofit		−0.46	−1.14	−0.19
f. Household Workers		−0.21	−0.21	−0.21
13. Government Enterprises	−0.51	1.10	−0.52	0.99
14. Government Industry	0.31	−0.18	0.08	−0.18
Overall: GDP	2.16			
GNP		2.18	1.17	2.18

Sources: Col. 1: BEA, *The National Income and Product Accounts of the United States*, 1929–76 Statistical Tables, September 1981, Tables 6.2 and 6.11. Col. 2: *GDO* for 1947 was obtained from the standard 87-order BEA input-output table for 1947; *GDO* for 1976 was obtained from BLS, *Time-Series Data for Input-Output Industries*, Bulletin 2018, 1979. Cols. 3–4: U.S. input-output data. See fn. 4 for details.

[a]Shown in percent.

estimated rates of sectoral productivity growth differ somewhat from those in column 1, though the rank orders are quite close. The major exception is the construction sector, whose 1.19 percent rate now places it in the stagnant category. The input-output data also permit disaggregation of general services into six subsectors, as shown in Table 1, and evaluation of their degrees of stagnancy. The range of sectoral productivity growth rates of these subsectors is fairly wide, though they all lie below the economy's 2.18 percent rate. The last three subsectors in this group all seem clearly to be stagnant. The first three are more marginal, though we will, somewhat arbitrarily, draw the line between business and professional services (1.70 percent) on the one hand, and hotels, personal and repair services (1.37) and auto services (1.45 percent) on the other, placing only the former in the progressive group.

Our third measure of productivity growth rates requires several symbols to describe the input-output framework. Let X = (column) vector of gross output by sector; Y = (column) vector of final demand by sector; a = matrix of interindustry technical coefficients; l = (row) vector of labor coefficients; k = (row) vector of capital stock coefficients; and p = (row) vector of prices showing the (current) price per unit of output of each industry. In addition, we use the following scalars: w = the annual wage rate, in current dollars; r = the rate of profit on the capital stock; $y = pY = GNP$ at current prices; $L = lX$ = total employment; and $K = kX$ = total capital stock.

The aggregate rate of total factor productivity (TFP) growth is given by

$$(1) \qquad \rho = (pdY - wdL - rdK)/y,$$

where d refers to the differential. The rate of TFP growth for sector j is given by

$$(2) \quad \rho_j \equiv -\Big(\sum_i p_i da_{ij} + wdl_j + rdk_j\Big)/p_i.$$

This is the continuous analog of Wassily Leontief's 1953 measure of sectoral technical change.[3]

The U.S. input-output data for 1947 and 1976 were used to estimate this third set of growth rates (col. 3, Table 1).[4] The TFP measures were all lower than the corresponding labor productivity measures since capital-labor ratios were increasing. The overall rate of TFP growth was 1.17 percent per year, about one point lower than that of labor productivity, and the sectoral rates behaved similarly. Their relative magnitudes were virtually unchanged, except for mining.[5] By this measure, the line between the progressive and stagnant categories was drawn between nondurable manufacturing (0.41 percent), and government industry and the mining sector (both at 0.08 percent).

So far, our productivity measures evaluate productivity improvements within any one sector; one can also examine the changes in total input usage, direct and indirect, per unit of a sector's output. This also reflects productivity growth of the sector's input suppliers. One such total factor requirement measure (reported in col. 4, Table 1) is λ, which shows the total (direct plus indirect) labor requirements per unit of final output:

$$(3) \qquad \lambda = l(I-a)^{-1}.$$

Productivity growth based on λ is quite similar to the figures in column 2, Table 1, since changes in total factor requirements are dominated by those in direct factor requirements.[6] The classification of sectors uses cut-

[3] Also, see William Peterson (1979) and Wolff (forthcoming) for more details. Because discrete time periods are employed, a Turnquist-Divisia Index is used to estimate sectoral and the overall rate of TFP (see Frank Gollop and Dale Jorgensen, 1980, or Wolff, forthcoming).

[4] The 1947 input-output table is the standard 87-order Bureau of Economic Analysis (BEA) version. (See, for example, BEA *Survey of Current Business*, 1974, for methods and a listing of sectors.) The 1976 table was estimated using the so-called R.A.S. method on the 1972 table, with the gross domestic output figures in Bureau of Labor Statistics (BLS, 1979a). Estimates of the total capital stock in each input-output sector appear in BLS (1979b). Full capital coefficient matrices for 1947 were obtained from the Brandeis Economic Research Center (BERC); sectoral 1947 depreciation rates from BERC; and those for 1976 estimated from *Internal Revenue Service Corporation Tax Returns*. Sectoral price indices for 1947 were provided by BERC and for 1976 by the BEA. Additional details on data sources and methods are available from the authors.

The accounting framework was then modified as follows: 1) An "endogeneous export column" was created to balance the noncompetitive import row (sector 80). 2) For the estimation of Marxian labor values, the depreciation row that is normally part of value-added was treated as an endogenous input row (sector 88), and an "endogenous capital replacement" column was included to balance this row. 3) Five sectors (research and development (74), business travel (81), office supplies (82), scrap and used goods (83), and inventory valuation adjustment (87)) appeared in the 1947 table but not in the 1976 table. In order to assure consistency of the accounting framework, these sectors were eliminated from both gross and final output in 1947 by distributing their inputs to other sectors. 4) Indirect business taxes in value-added were eliminated in order to remove the biasing effect of indirect business taxes on relative prices. 5) The input-output matrices were finally converted to constant (1958) prices by multiplying each row of the matrix by the appropriate sectoral price deflator. For details, see Wolff (forthcoming).

[5] This reflects a large postwar influx of capital equipment into mining and increases in intermediate inputs. The mining sector is rather different from a more standard stagnant sector, since it is a process industry whose output is not directly related to its labor (or capital) input. Its low rate of TFP growth is attributable primarily to the nature of extraction, in which more accessible ores and petroleum are mined first and less accessible deposits later. The increasing difficulty of mining would have yielded a negative growth rate in TFP if technology had remained constant. The fact that TFP growth was zero in this sector over the period 1947–76 suggests that technical change (or the discovery of new accessible deposits) did occur.

[6] Three other measures were also used. The first, λ_m, differs from col. 4 only in λ's Marxian accounting framework. Capital, as a produced means of production, is valued by its depreciation rate (see Wolff, 1979). The second is ρ^*, the total factor requirement analog of ρ: Let $\gamma = k(I-a)^{-1}$ be the total capital requirements per unit of final output. Then the rate of change of total factor requirements per unit of final output can be estimated from $p_j^* \equiv -(wd\lambda_j + rd\gamma_j)/p_j$. The third

TABLE 2—SHARE OF EMPLOYMENT AND OUTPUT IN STAGNANT SECTOR, 1947 AND 1976[a]

	Measure			
	GPO/L (1)	*GDO/L* (2)	ρ (3)	λ (4)
A. Stagnant Sectors:				
2. Mining			×	×
3. Construction		×	×	×
10. Finance and Insurance	×	×	×	×
12. General and Services				
a. Hotels, Personal and Repair (except auto)	×	×	×	×
b. Business and Professional	×			
c. Auto Repair and Service	×	×	×	×
d. Movies and Amusement	×	×	×	×
e. Medical, Educational and Nonprofit	×	×	×	×
f. Household Workers	×	×	×	×
13. Government Enterprises	×	×	×	×
14. Government Industry	×	×	×	×
B. Annual Prod. Growth Rate, 1947–76:				
a. Progressive Sectors (all)	2.94	3.04	1.09	2.92
b. Stagnant Sectors	0.64	0.56	−0.84	0.73
c. Progressive Service Sectors	2.71	2.79	1.63	2.79
d. Overall	2.16	2.18	1.17	2.18
C. Percent of Employed Persons in Stagnant Sectors:				
a. 1947	27.6	30.7	32.4	32.4
b. 1976	41.2	42.0	43.0	43.0
D. Stagnant Sector Share of Final Output (1958 $):				
a. 1947	21.4	31.2	31.5	31.5
b. 1976	21.2	29.2	28.9	28.9
E. Stagnant Sector Share of Final Output (Current $):				
a. 1947	17.9	26.8	27.0	27.0
b. 1976	29.9	38.6	38.1	38.1
F. Stagnant Sector Share of *GDO* (1958 $):				
a. 1947	16.8	21.9	24.2	24.2
b. 1976	16.8	19.8	21.3	21.3
G. Stagnant Sector Share of *GDO* (Current $):				
a. 1947	13.7	18.3	20.4	20.4
b. 1976	22.9	24.5	26.7	26.7
H. Percent of Employed Persons in Progressive Services:[b]				
a. 1947	21.3	23.5	23.5	23.5
b. 1976	22.5	26.7	26.7	26.7

[a]Panels B–H results are shown in percent.
[b]In col. 1, progressive services are defined as communications and broadcasting, trade, and real estate. In cols. 2–4, they include the same three sectors and, in addition, business and professional services.

off points of 2.19 percent for the progressive category and 1.49 percent for the stagnant category, and is identical with that of column 2, except that the mining sector now falls into the stagnant category.

In Table 2, × indicates that a sector is classified as stagnant according to the measure of productivity growth (panel A). The average annual rate of productivity growth for the two aggregated sectors are shown in panel B.[7]

measure uses the rate of change in the (real) relative price of a sector's output to measure its relative rate of productivity growth. All three measures yielded the same classification scheme as shown in cols. 3 and 4 of Tables 1 and 2.

[7]It should be noted that the overall level of productivity growth corresponding to λ_m is the ratio of *NNP* to employment, since depreciation is treated as endogenous. The rate of growth is lower than that of *GNP* per worker.

V. Tests of the Model's Basic Implications

We are now in a position to test as hypotheses the main implications of our model. The first of these is the cost disease prediction that relative prices of the stagnant sector's outputs will rise at about the same rate as the shortfall in its rate of productivity growth. This is indeed confirmed by the data. By the measures of Table 2, the rate of productivity growth of the stagnant sector is from about two to two and one-half percentage points below that of the progressive sector.[8] Independently selected price data show that the price of stagnant output relative to progressive output increased at about 2 percent per year.

The next hypothesis is not an implication of the model, but was previously only a casual observation. This is the view that in real terms output shares have remained constant over time. This was examined using both final output and *GDO* shares (panels D and F, Table 2). The classification scheme of column 1 tells us that the real output shares remained constant over the period in terms of both final output and *GDO*. The other definitions, however, indicate a slight decline in the stagnant sector's real share of final demand and gross output.

We can now examine the other two main implications of the model. The first is that, since output shares have been fairly constant, the share of employment in the stagnant sector will rise over time. By all four definitions, the share of employment in the stagnant sector rose by over ten percentage points over the period and, by the first definition, by almost fourteen percentage points (panel C). The third basic prediction of our model is that, with output shares roughly constant in real terms, the share of output produced by the stagnant sector will rise in nominal terms over time. This is confirmed in panels E and G, which exhibit increases that range from 6 to 12 percent.[9]

One final set of implications of the model can also be tested. As has been shown, the service sector includes both progressive and stagnant industries. In panel B, we have calculated separately the rate of productivity growth for progressive services. We find that the progressive services experienced slightly lower rates of growth of labor productivity than progressive goods producers but higher rates of total factor productivity growth. Moreover (panel H), we find that while employment in progressive services increased over the 1947–76 period, it rose very modestly, as our analysis might lead us to expect. Thus, progressive services behaved very differently from stagnant services over the postwar period and behaved very much like progressive goods sectors, and while it is true that the nation's labor force moved toward services, both stagnant and progressive, it was the former whose labor force increased most substantially. While the labor force of the progressive services rose somewhere between 5 and 14 percent, that of the stagnant services rose between 32 and 50 percent.[10]

[8]Both sectoral values of ρ are below the overall rate of *TFP* growth. This is correct, because as demonstrated in Peterson, $\rho=\Sigma_i(p_iX_i/y)\rho_i$, the ratio of total *GDO* to total final output (in current dollars) is about 2.0 in both years.

[9]We also found that the share of total capital stock in the stagnant sector declined by about five percentage points, indicating that the capital-labor ratio grew faster in the progressive sector. This result is consistent with the spirit of our model, since the progressive sector is characterized by more rapid changes in technology that can be expected to involve a more rapid displacement of labor by capital.

[10]Some remarkable cross-sectional international comparisons provided by Robert Summers (1985) also offer

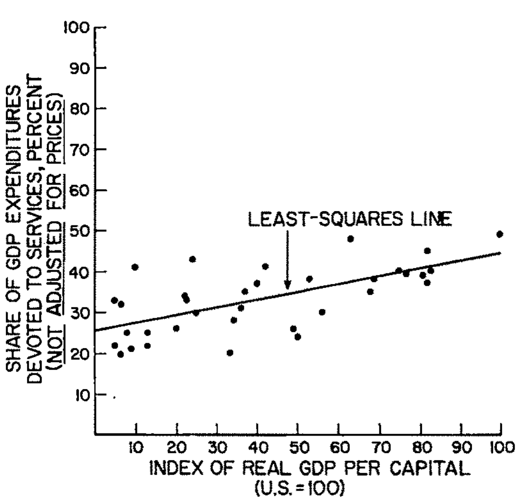


FIGURE F1

VI. Broadcasting, Electronic Computation, and Asymptotic Stagnancy

Our empirical evidence on two asymptotically stagnant activities, television broadcasting and data processing (computer services), shows that in both activities the progressive component's share of total costs diminished continually, while the stagnant component increased both in real terms and as a share of total cost.

A. *Electronic Computation*

In the last twenty years the cost of computer hardware per unit of processing power apparently fell some 25 percent per year (see, for example, W. J. Kubitz, 1980; S. Triebwasser, 1978; R. N. Noyce, 1977; and C. Burns, 1977). Meanwhile, the cost of (labor-intensive) computer software assumed an ever greater share of a computer system's total cost. Software was once a relatively minor element in computing cost—indeed, IBM once gave software away with its machines. Now, it is the hardware that is becoming almost incidental in total computation cost (see T. J. Gordon and T. R. Munson, 1980). By some estimates, software represented only 5 percent of system costs in 1973, had increased to 80 percent by 1978, and exceeded 90 percent by 1980 (see Kubitz; M. Schindler, 1979; and R. A. Minicucci, 1982). P. Grabscheid writes, that by 1985, "it will probably pay to substitute one hour of computer time for six minutes of staff time" (1982, p. 6). Software development remains essentially a handicraft activity, and is, so far, a stagnant service.

Some operating data from the Princeton University Computer Center (Figure 1) substantiate dramatically the growing importance of labor costs in total Center expenditures and the accompanying sharp drop in the dominance of the hardware component.[11] Between 1970 and 1983, total real labor costs at the Center rose at a compound rate of 2.6 per annum, while total real equipment costs fell at an annual rate of 4.6 percent.[12] Since the volume of computations has risen rapidly,

at least suggestive support for our model. The services' proportion of total real *GDP* expenditures, and their proportion of total nominal *GDP* expenditures were compared with real *GDP* per capita for a sample of 34 countries, ranging from very poor countries such as Malawi and India to highly industrialized states like Germany and the United States. As Figure F1 reports, at least in 1975, the real share of the services did not increase with a country's real per capita *GDP*, contrary to widespread belief. However, as our model suggests, since the real share of *GDP* devoted to services remained roughly constant among countries, the nominal share devoted to services nevertheless rose markedly with real *GDP* per capita (Figure F2). The results of a regression were completely consistent with these conclusions.

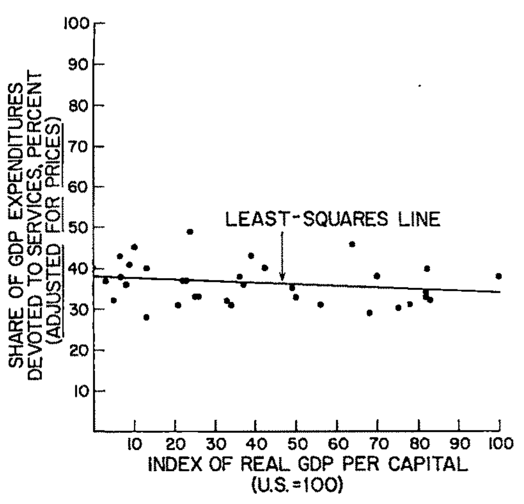


FIGURE F2

[11] In the three years (1976; 1979; 1981) in which the downward trend was interrupted, the increased share of hardware cost is ascribable to major equipment purchases and changes in equipment financing, rather than to increases in hardware prices. The Director of the Center does caution that, although the bulk of the drop in Center expenditures on hardware is attributable to actual hardware cost decreases, some part of it is the result of more favorable lease-purchase arrangements and an increase in the percent of equipment owned rather than rented.

[12] Some industry figures produce results that are less clear cut. For instance, the Diebold Group (1982) has studied computer operations of large U.S. corporations over the ten-year period, 1971–81. Their surveys showed that the average share of computer operations budgets devoted to hardware fell from 35 percent in 1971 to 27 percent in 1981; the share of expenditures on operations personnel (i.e., keypunchers whose work is most susceptible to automation and productivity increases) fell from 29 percent in 1971 to 18 percent in 1981; while the share of the budget spent on systems development personnel (the "brainpower" employees) remained essentially the same over the ten-year period (25 percent in 1971 and 24 percent in 1981).

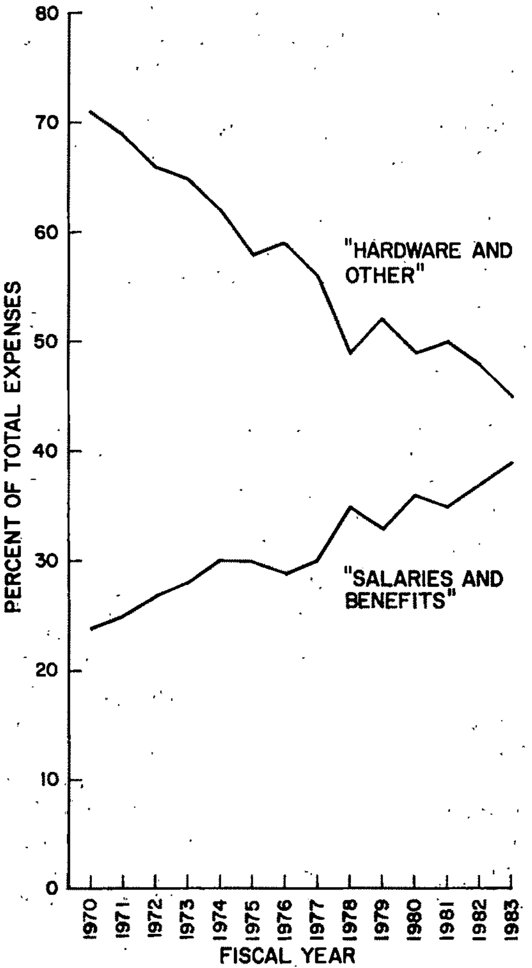


FIGURE 1. LABOR COSTS VS. HARDWARE COSTS AS A PERCENTAGE OF TOTAL COSTS, PRINCETON UNIVERSITY COMPUTER CENTER, 1970–83

Source: James Poage, Director, Princeton University Computer Center.

Notes: The cost category "Hardware and Other" is made up of approximately 80 percent computer hardware costs and 20 percent other costs, such as disposable supplies. Increases in hardware costs in 1976, 1979, and 1981 are largely ascribable to either the purchase of major new equipment, or the refinancing of equipment costs. Staff size at the Center has remained essentially unchanged over the period. Data for 1983 are estimates.

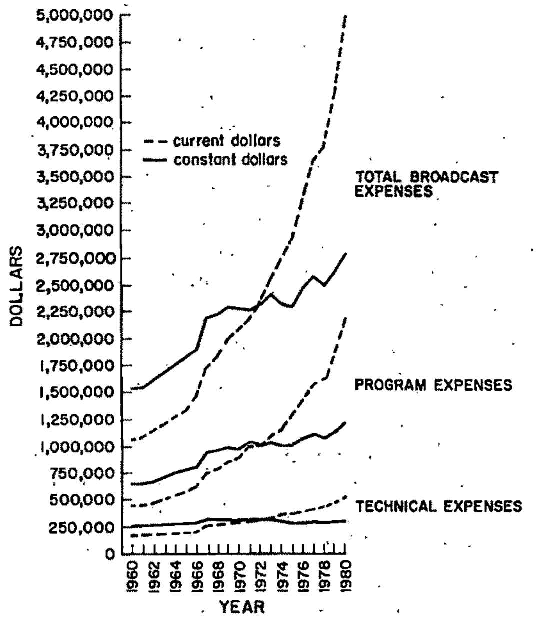


FIGURE 2. BROADCASTING EXPENSES PER AVERAGE TELEVISION STATION[a]

[a]Shown in current and constant dollars. Data exclude the three major networks, but include network owned and operated television stations.

Sources: U.S. Federal Communications Commission, *Annual Report*, various years, and "Television Financial Data 1980, FCC Financial Figures," August 10, 1981, No. 6, Vol. 101, p. 54. Source for price deflator is *Survey of Current Business*, various years.

equipment cost per unit of output has fallen far more rapidly (and per unit labor costs have risen more slowly).[13]

[13]Although the number of computations performed at the Center is not recorded, according to the Director of the Center, this number has clearly increased dramatically. In particular, as the computer programs handled at the Center became ever more complex (i.e., as the "captured intelligence" in each program grew), each keystroke punched into the computer gave many more commands to the machine. We should note here that the other side of the phenomenon of the increasing domination of labor costs in computer budgets is the extraordinary increase in labor productivity brought about by computerization. Computer technology permits users to accomplish much more much faster. For example, a company that once paid a roomful of workers to tabulate year-end accounts can now computerize those operations and retrain the workers to analyze the data the computer puts out. At the Princeton University Computer Center the budget for salaries used to be dominated by keypunch personnel; today the staff there is far more skilled and professional. The data processing industry is seeking ways to enhance further the productivity of its personnel, for example, by finding ways to substitute hardware time for costly staff time and by creating software in so-called "fourth generation" computer languages which minimize the user's time and permit less-skilled (and lower-paid) operators to use the computer.

TABLE 3—TECHNICAL AND PROGRAM EXPENSES AS A PERCENTAGE OF TOTAL TELEVISION BROADCASTING EXPENSES, 1960–80

Year	Total Broadcast Expenses (all TV stations,[a] in millions of dollars)	Technical Expenses (mil. of $)	Technical Expenses as Percent of Total	Program Expenses (mil. of $)	Program Expenses as Percent of Total
1960	563.3	92.9	16.5	239.1	42.4
1961	579.5	96.2	16.6	245.2	42.3
1962	626.6	101.3	16.1	265.4	42.3
1963	674.5	106.3	15.8	290.5	43.1
1964	725.4	113.8	15.7	315.1	43.4
1965	787.7	120.2	15.3	338.8	43.0
1966	885.0	131.1	14.8	380.1	43.0
1967	948.3	141.4	14.9	409.2	43.2
1968	1040.1	151.5	14.6	449.2	43.2
1969	1176.4	161.4	13.7	504.9	42.9
1970	1245.2	170.6	13.7	534.7	43.0
1971	1303.7	179.6	13.8	599.2	46.0
1972	1457.6	196.5	13.5	628.6	43.0
1973	1577.9	210.5	13.3	677.9	43.0
1974	1706.7	228.3	13.4	733.7	43.0
1975	1830.0	229.6	12.5	805.3	44.0
1976	2108.1	256.7	12.2	912.3	43.3
1977	2297.1	270.3	11.8	995.1	43.3
1978	2705.4	318.4	11.8	1162.5	43.0
1979	3100.6	346.3	11.2	1343.6	43.3
1980	3614.6	390.0	10.8	1588.3	43.9

Source: U.S. Federal Communications Commission, *Annual Report*, various years.
Notes: Technical expenses include payroll and other technical expenses such as circuit costs incurred in delivering programs to local stations. Program expenses include "talent" employees, other employees, rent and amortization of film and tape, records and transcripts, outside news service costs, payment to talent, music license fees, other performance and program rights, and all other program expenses. Other categories not listed in the table are selling expenses and general and administrative expenses (which includes general and administrative payroll, depreciation and amortization, interest, allocated costs of management from home office of affiliates(s), and other general and administrative expenses). These descriptions are taken from "Television Financial Data 1980, FCC Financial Figures," *Broadcasting*, August 10, 1981, Vol. 101, No. 6.
[a]Does not include the three major television networks but does include network owned and operated television stations.

B. *Television Broadcasting*

Television broadcasting also has progressive and stagnant components, such as transmission, which includes circuit costs, and programming, dominated by human labor. Here, too, the evidence on trends in costs, and trends in cost shares, is striking. Figure 2, using U.S. Federal Communications Commission data, shows the steep rise in average expenses of TV stations between 1960 and 1980 (in both current and constant dollars), and portrays the trends in the two relevant components of broadcasting expenses (technical and program expenses), showing that real program costs have climbed steadily, while real technical expenses have remained about constant over the twenty-year period. In Table 3 we see that, as a percent of total expenditures, technical costs have dropped continuously from 16.5 percent in 1960 to 10.8 percent in 1980. In constant dollars, over the twenty years in question total technical expenses per station have actually risen, but at the modest rate of 0.8 percent per year. However, the average rate of increase of real programming cost was 3.1 percent, and total real expenses increased at virtually the same annual rate, 2.9 percent.

VI. Concluding Comments

All the empirical data we have found seem consistent with the predictions of the amended unbalanced growth model. The

"rising share of services" turns out to be somewhat illusory. The output shares of the progressive and stagnant sectors have in fact remained fairly constant in the postwar period, so that with rising relative prices, the share of total expenditures on the (stagnant) services and their share of the labor force have risen dramatically (their prices rose at about the same rate as their productivity lagged behind the progressive sectors), just as the model suggests. Similar trends are also found internationally.

We have also introduced into the model a type of activity we call asymptotically stagnant—economic enterprises which seem among the most high tech and progressive one can imagine. They contain both a technologically sophisticated component and a relatively irreducible labor-intensive component. Starting out as innovative activities dominated by their very productive technological side, as the labor component assumes an ever larger share of total cost (because the progressive component is innovating itself out of its cost-dominating position), ultimately the activity assumes all the characteristics of the stagnant services. Empirical data on two such activities—TV broadcasting and electronic computation—are also consistent with the model's predictions. This suggests that the progressivity of such activities may well prove transitory and somewhat illusory. In sum, the cost disease of the stagnant services may affect more of the economy than was previously thought.

REFERENCES

Baumol, W. J., "Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis," *American Economic Review*, June 1967, *57*, 415–26.

Burns, C., "The Evolution of Office Information Systems," *Datamation*, No. 4, 1977, *23*, 60–64.

Gollop, Frank M. and Jorgensen, Dale W., "U.S. Productivity Growth by Industry, 1947–73," in John W. Kendrick and Beatrice N. Vaccara, eds., *New Developments in Productivity Measurement and Analysis*, Chicago: University of Chicago Press, 1980.

Gordon, T. J. and Munson, T. R., "Research Into Technology Output Measures," unpublished paper, The Futures Group for the National Science Foundation, November 1980.

Grabscheid, P., "The Economics of Information Processing," presentation for 1982–83 Chief Financial Officer Seminar Program Series, *Institutional Investor*, 1982.

Kubitz, W. J., "Computer Technology, A Forecast for the Future," in F. Wilfrid Lancaster, ed., *Proceedings of the 1979 Clinic on Library Applications of Data Processing, The Role of the Library in an Electronic Society*, Urbana-Champaign: University of Illinois Graduate School of Library Science, 1980, 135–61.

Leontief, Wassily, *Studies in the Structure of the American Economy*, New York: Oxford University Press, 1953.

Levin, H. J. *Fact and Fancy in Television Regulation*, New York: Russell Sage Foundation, 1980 (who cites Federal Communications Commission Network Inquiry Special Staff, *The Historical Evolution of the Commercial Network Broadcast System*, October 1979, p. 176).

Minicucci, R. A., "Sub-second Response Time: A Way to Improve Interactive User Productivity," *Systems Management Controls, SMC Newsletter* 82–19, November 1982.

Noyce, R. N., "Microelectronics," *Scientific American*, No. 3, 1977, *237*, 63–69.

Paik, N. J., "How to Keep Experimental Video on PBS National Programming," in *Independent Television-Makers and Public Communications Policy*, Rockefeller Foundation Working Papers, December 1979, ch. 2.

Peterson, William, "Total Factor Productivity in the U.K.: A Disaggregated Analysis," in K. D. Patterson and Kerry Scott, eds., *The Measurement of Capital: Theory and Practice*, London: Macmillan 1979.

Schindler, M., "Computers, Big and Small, Still Spreading as Software Grows," *Electronic Design*, No. 1, 1979, *27*, 88.

Summers, Robert, "Services in the International Economy," ARA/Wharton Conference on the Future of the Service Economy, University of Pennsylvania, Philadelphia, forthcoming 1985.

Triebwasser, S., "Impact of Semiconductor Microelectronics," *Computer Technology: Status, Limits, Alternatives*, New York: Institute of Electrical and Electronics Engineers, Inc., 1978, 176–77.

Wolff, Edward N., "The Rate of Surplus Value, the Organic Composition, and the General Rate of Profit in the U.S. Economy, 1947–1967," *American Economic Review*, June 1979, *69*, 329–41.

______, "Industrial Composition, Interindustry Effects, and the U.S. Productivity Slowdown," *Review of Economics and Statistics*, forthcoming.

The Diebold Group, "Management Information Services/Telecommunications Budgets, 1982," Document Number 211M, Abstract, p. 10 (also personal communication with David Dell, Director of Research Services, the Diebold Group, Inc., New York, NY).

U.S. Federal Communications Commission, *Annual Report*, various years.

U.S. Department of Commerce, Bureau of Economic Analysis, "The Input-Output Structure of the U.S. Economy: 1967," *Survey of Current Business*, February 1974; and various years.

______, ______, *The National Income and Products Accounts of the United States, 1929–1976*, Statistical Tables, Washington, 1981.

______, **Bureau of Labor Statistics,** (1979a) *Time-Series Data for Input-Output Industries*, Bulletin 2018, Washington, 1979.

______, ______, (1979b) *Capital Stock Estimates for Input-Output Industries: Methods and Data*, Bulletin 203, Washington, 1979.

Labor Market Costs of Language Disparity: An Interpretation of Hispanic Earnings Differences

By WALTER S. MCMANUS*

Persons with limited English proficiency are a significant and growing minority in the United States. In 1980 over 23 million persons (11 percent of the U.S. population over four-years old) reported that English was not their only language of communication.[1] Of these, 45 percent, or about 5 percent of the total, also reported a less than fluent command of the English language. To the extent that English is, and will remain, the predominant language of interpersonal communication in the United States, the potential exists for labor market earnings to be reduced by a lack of proficiency in English. Further, to the extent that earnings reflect marginal value productivities of workers, the maintenance of minority languages by a growing segment of the labor force will result in significant social costs, in terms of a reduced value of output.

Among minority languages in the United states, Spanish is the most common. In 1980, Spanish speakers accounted for 46 percent of the language minority population. Not all Hispanics are Spanish speakers. In 1976, about 18 percent of the male Hispanic labor force were English monolinguals. However, Spanish heritage plays a crucial role in defining the Hispanic ethnic identity. After blacks, Hispanics are the largest ethnic minority in the United States. According to the 1980 census, Hispanics account for 6.4 percent of all U.S. residents, while blacks account for 11.7 percent. The growing importance of this language-associated minority warrants analytical attention.

A small but growing body of literature has begun the analysis of the effects of language skills on earnings of Hispanics. Using detailed earnings information from the *Survey of Income and Education* (*SIE*), Cordelia Reimers (1983) constructed a simple measure of English ability, but did not find significant language effects except among Puerto Ricans and "other" Hispanics. Gilles Grenier (1984) goes beyond Reimers by using more of the available information on language characteristics in the *SIE* data. He found that language attributes play an important role in wage determination, explaining up to one-third of the relative wage difference between Anglo and Hispanic men. My article with William Gould and Finis Welch (1983), also using the *SIE*, found that virtually all the Hispanic-Anglo wage differences usually attributed to ethnicity, nativity, and time in the United States can be explained by differences associated with English language skills. This empirical finding is the starting point for the current analysis concentrating on differentials within the Hispanic group.

In Table 1, I present the average relative wages of Hispanic men by English proficiency. Proficiency in English is determined by responses to questions of how well English is understood and spoken. These questions are only asked of persons who report frequently using a language other than English. What I call the Fluent Reference Group is comprised of Hispanic men not asked the English proficiency questions. Wages are reported relative to the wages of the Fluent Reference Group. Table 1 demonstrates the gross correlation between earnings and English proficiency; the greater the English proficiency, the higher the relative wage.

My plan is to first review the theoretical issues involved by considering several models of language disparity in markets, and then to present an empirical analysis of Hispanic wage differences in the United States for men 16–65-years of age in 1975. The goal of the empirical analysis is to estimate the cost

*Assistant Professor of Economics, University of Florida, Gainesville, FL 32611. I have benefited from conversations with Finis Welch and participants in the University of Florida's Labor Workshop.

[1] *1980 Census of Population.*

TABLE 1–WEEKLY WAGES OF HISPANIC MEN BY ENGLISH PROFICIENCY AS A PERCENTAGE OF WEEKLY WAGES OF ENGLISH-FLUENT HISPANIC MEN, 1975

English Proficiency Group[a]	Relative Weekly Wage
Fluent Reference Group[b]	100.0
Very Well	89.8
Well	82.8
Not Well	68.2

Source: U.S. Department of Commerce, Bureau of Census, *Survey of Income and Education* (*SIE*) microdata file (computer tape), 1976.

[a]English Proficiency Group is defined by responses to questions: "How well does ______ understand/speak English?"

[b]The Fluent Reference Group includes Hispanic men not asked how well they understand/speak English. They are assumed fluent, based on their background.

of language disparity for Spanish-speaking men, by identifying wage differences that can be attributed to a lack of English proficiency. Using this procedure, I estimate that the cost of language disparity for Hispanics ranges between $1000 and $2000 per year. The present value of acquiring perfect fluency in English is between $19,000 and $36,000 for the Hispanic men in my sample. These estimated costs of friction are the potential gains to language-training programs, and should be compared to language-training costs to determine whether these programs could have net benefits.

I. Theoretical Background[2]

In an economy where English is the dominant language of communication, it seems obvious that an inability to communicate in English has the potential for lowering earnings. The extent of the earnings loss depends both on the extent to which production is integrated and on the nature of the product. For impersonal products, any discount of value by the consumer for not being able to communicate with the producer must be caused at the time of purchase, for thereafter communication is irrelevant. For standardized products, not only is the identity of the producer irrelevant, but also consumption does not intimately involve communication. Bilingual salesmen of a standardized product have a comparative advantage, in a heterogeneous environment, over salesmen with only one language. It is also likely that the degree of proficiency in a language needed to act as an intermediary for impersonal products is less than is needed for personal products. Thus, the supply of potential intermediaries is relatively larger for standarized than for personalized products, so price differentials are likely to be small.

For personal products, the identity of the producer is relevant, so that the value of the service to consumers may depend on the producer's ability to communicate with them. Differences in price between language groups can emerge due to differences in demand and supply conditions. These differences can also lead to premiums for knowledge of English among minority language producers for markets dominated by English-speaking consumers. If an individual supplier could use knowledge of two languages to serve as a translator-intermediary for personalized products, there would potentially be a premium for bilinguality. However, the personalized nature of the product would severely limit intermediation between monolinguals from different groups.

In production, the languages of workers, managers, and capital are significant only if there is an incentive to integrate, and only if integration is enhanced by inputs being able to communicate with each other. If an industry were composed solely of one-person-producing units, if the production were static and involved only labor, then communication skills would not affect income. If each person produced only corn or wheat, then trade would establish a wheat price in terms of corn, and each producer's earnings would reflect only his physical production and the market price.

If production technology is not static, but continually improving, then even one-person-producing units should benefit by being able to use new technology. The ability to absorb new techniques depends, in turn, on the ability to communicate with other producing units and/or with the agency or per-

[2]See my 1983 dissertation for more discussion of theoretical issues.

son who develops the new technique. Hence, if technical advances in production are more likely to be made by members of one language group, then the earnings of members of other language groups ought to be enhanced by also speaking the technologically advancing language. Not only could a bilingual use new technical advances in both of his languages, but he could also serve as a translator of technology from one to the other. Here the ability to speak modal language enhances earnings only if it is the modal language group that makes the most technical advances. If there are economies of scale in technical progress and experimentation, then I would expect the modal language group to be the source of a disproportionate share of advances in production technology.

Similarly, if production is integrated into multiperson units, then the complementarity or jointness that draws workers together should be enhanced by their abilities to communicate with each other. If members of a team cannot communicate with each other, they may be unable to fully realize the benefits from specialization of function by different groups. The incentive to integrate into multiperson units is derived from assuming that persons differ in their endowments of some productive inputs, say education, that makes workers with differing amounts of educations complements in production. In turn, any complementarity must be the result of specialization of function by the workers. It is therefore possible that the full amount of the complementarities that would be derived from combining two laborers from the same language group is not derived when the laborers of two language groups are integrated. The information exchange necessary for the specialization cannot take place.

It is also possible that intergroup association could result in overspecialization of labor. Combining laborers of two language groups might result in a precise division of tasks based on average characteristics of groups with no allowance for personal variation. Division of tasks might be constrained so that a given task must always be performed by a member of group 1 even in circumstances where a member of group 2 could perform the task with less resource cost. This case would be likely if one group were unable to communicate to management their ability to perform certain nonmenial tasks. I would expect that speakers of nonmodal languages would be likely to be relegated to menial tasks, the requirements of which could be easily communicated by demonstration without the need for verbal explanations.

II. Measuring Costs Of Language Disparity

In this section I estimate costs of language disparity for Hispanic men in the United States in 1975. Data are from the 1976 *SIE*. This is a census-style survey that contains information on labor force status, work experience, income in 1975; along with basic demographic items such as age, race, sex, schooling attainment, and ethnic origin. Responses to these questions provide the background for the analysis. The *SIE* also includes information on detailed ethnicity for Hispanics, nativity and year of immigration to the United States for immigrants, years of schooling prior to immigration, and the number of years of this pre-immigration schooling that was in English. Finally, the *SIE* provides a wealth of information on language characteristics: 1) household languages spoken; 2) childhood language; 3) personal languages; 4) languages with friends and children; and 5) proficiency in understanding and speaking English. I used responses to these questions to develop measures of English proficiency and to identify bilingual Hispanics.[3]

The sample that I used for analysis includes information on 69,844 native-born English-speaking white non-Hispanic men (referred to as "Anglo" in this paper), 5,389 native-born English-speaking black non-Hispanic men, and 2,576 Hispanic men. The samples all refer to civilians ages 16–64, not now in school (as major activity last week), who either worked 50 to 52 weeks in 1975, or report the reason for fewer weeks as some-

[3]For a detailed description of the *SIE*, see my dissertation.

thing other than retirement or school. Those self-employed or working without pay are excluded. In addition, those with zero or missing information whether of weeks worked or of wage and salary earnings in 1975, and those whose weekly wage (computed as the ratio of wage and salary earnings to weeks worked) is less than $10 are also excluded. Hispanic men were also excluded from the analysis if 1) they had missing or inconsistent nativity or language information, 2) they immigrated in 1975 or 1976, 3) they reported using any language other than English or Spanish, or 4) they reported pathological language information.[4]

In the earnings analysis reported earlier in my dissertation, measures of language skills were used in fairly standard wage regressions for Hispanic men. Here I will use these estimated effects of language skills to measure the costs of language disparity.

A. *Measuring English Proficiency*

The *SIE* does not give a single measure of language skills, but rather gives responses to a series of questions concerning languages used in the home, personal languages, languages with friends and children, newspaper reading, and levels of understanding of spoken English. In principle, these questions could divide Hispanic men into over 20,000 groups. Reducing the size of this problem to manageable proportions was the bulk of the study reported in my article with Gould and Welch. There three sets of questions were used to form language groups: household language; personal language; and self-reported English ability. These three sets of questions split Hispanic men into just 42 groups. These 42 groups were then aggregated up into 7 groups that captured the bulk of the language questions of the *SIE*, and at the same time were arguably well-ordered in terms of English proficiency. Membership in a particular aggregate was largely determined by household language with personal language and proficiency playing marginal roles.

An attempt to reevaluate the use of the *SIE* questions to construct indicators of English proficiency and bilinguality was motivated by several considerations. First, the dependence of the aggregation scheme of my article with Gould and Welch on household languages raises serious questions of whether what is being measured is personal English proficiency or cultural assimilation. Household language is more likely to be contaminated with cultural aspects of assimilation than personal language. Second, the aggregation scheme did not attempt to distinguish between monolingual and bilingual English fluent men. In addition, the use of all the questions to construct *one* indicator makes comparison with other data sets problematical. Simple indicators based on only a few related questions are more likely to be obtainable for other data sets. Comparison of results using different indicators within one data set may also be enlightening.

I have dropped household language as an indicator of personal English proficiency. The question of how a household language is chosen is deferred to future research. For now it is sufficient to remark that it is clearly affected by characteristics of household membership that are not normally considered as determinants of personal labor market success. I am left with two sets of questions that potentially can serve as indicators of English proficiency: personal languages and self-reported English understanding and speaking ability. Both contain only a few well-ordered response categories and are likely to be repeated in future data sets.

Personal language groups are: 1) English only or not asked; 2) English first, Spanish second; 3) English second, Spanish first; and 4) Spanish only. Proficiency in understanding and speaking English had responses: 1) Not Asked; 2) Very Well; 3) Well; and 4) Not Well.[5] Rather than consolidate these

[4] For further information regarding my selection criteria, see my dissertation.

[5] "Not well—more than a few words," "Not well—a few words," and "Not at all" categories were consolidated based on the ranking of dummy coefficients in an earnings regression. Persons not asked concerning English proficiency were assumed to be fully fluent.

potential measures into only one indicator, I compared results in my dissertation using either. Here I report results using self-reported proficiency in understanding and speaking English. The results using the other measure are essentially the same.

In the remaining earnings analysis, the four groups of the English proficiency indicator are collapsed into a scalar index of English proficiency (EP) by the convention:

$$EP = \delta_2 D_2 + \delta_3 D_3 + D_4, \tag{1}$$

where D_i is a (0,1) dummy indicating membership in the ith proficiency group and δ_i's are weights to be estimated. The D_i's are ordered so that the most proficient group is D_4. The origin ($\delta_1 = 0$) and scale ($\delta_4 = 1$) specify the convention under which EP is viewed as an index of proficiency.

The *SIE* has no direct measure of Spanish proficiency (SP) for assessing bilingual premia. If persons deficient in English are assumed to be fluent in Spanish, then all the men in EP groups 1 through 3 have $SP = 1$. Within EP group 4, I assigned a value of $SP = 1$ for men who had a Spanish childhood but who currently speak English only, and $SP = 0$ otherwise. A bilingual is thus a man who reports English as his only current language, but who is potentially fluent in Spanish due to his childhood ($EP = 1$, $SP = 1$).

B. *English Proficiency, Bilinguality, and Earnings*

Given measures of EP and SP, I turn now to the regression analysis of earnings of Hispanic men. I am interested in estimating effects of EP and SP on earnings as well as in how the estimates of effects of special variables associated with ethnicity and nativity are affected by the inclusion of EP and SP. To specify effects of EP, I assume that the income gain associated with higher skills in communicating in English would themselves be proportionally greater at higher levels of skill as measured by schooling and work experience. The theory section suggests that a potential source of this interaction arises from higher EP increasing the wage that labor skills can demand. Another explanation might be that lack of EP impairs acquisition of other skills relevant to an English-dominated labor market. Regardless of the source, there is a strong case for considering EP schooling and experience interactions. The EP variable was also interacted with residence not in an SMSA. Spanish proficiency was only entered directly, because an attempt at a richer specification did not yield significant results.

The model estimated is of the form,

$$\begin{aligned} &\mathit{Log(Wage)} = \mathit{common\ effects} \\ &+ \mathit{special\ effects} + EP^*\mathit{Delta} + c^*\mathit{Sp} + \mathit{error}, \end{aligned} \tag{2}$$

where *common effects* include effects of schooling, experience, and region of residence; *special effects* include Hispanic ethnicity, native birth, and foreign experience and schooling; EP is defined as in equation (1); SP is the indicator of Spanish proficiency; and

$$\begin{aligned} \mathit{Delta} = \mathrm{a}_0 + \mathrm{a}_1\ &\mathit{schooling} \\ &+ \mathrm{a}_2\ \mathit{experience} + \mathrm{a}_3 NS \end{aligned} \tag{3}$$

where NS is the indicator for not being an SMSA resident.

This model cannot be estimated by ordinary least squares (OLS), because nonlinear restrictions are imposed.[6] Instead, nonlinear least squares (the SAS procedure

[6] The EP effect is

$$EP^*\mathit{Delta} = (\delta_2 D_2 + \delta_3 D_3 + D_4)^*(a_0 + a_1 S + a_2 E + a_3 N)$$

where S is schooling, E is experience, and N is the non-SMSA indicator. Multiplying through gives

$$\begin{aligned} EP^*\mathit{Delta} = {} & a_0\delta_2 D_2 + a_0\delta_3 D_3 + a_0 D_4 + a_1\delta_2 D_2 S \\ & + a_1\delta_3 D_3 S + a_1 D_4 S + a_2\delta_2 D_2 E + a_2\delta_3 D_3 E \\ & + a_2 D_4 E + a_3\delta_2 D_2 N + a_3\delta_3 D_3 N + a_3 D_4 N. \end{aligned}$$

The constraints are of the form: the ratio of coefficients on D_2 and D_3 equals the ratio of coefficients on $D_2 S$ and $D_3 S$.

TABLE 2—NONLINEAR REGRESSION ESTIMATES OF EFFECTS OF LANGUAGE CHARACTERISTICS OF WEEKLY WAGES OF HISPANIC MEN, 1975[a]

	Parameter Estimate	Absolute t-Ratio
Common Variables:		
Constant[b]	3.600	13.97
Schooling 1 (0–12)	0.084	3.65
Schooling 2 (above 12)	0.059	3.30
Experience 1 (0–15)	0.082	4.56
Experience 2 (above 15)	0.029	2.64
Experience Squared	−0.0003	3.00
Schooling 1, Experience 1 Interaction	−0.003	1.88
Schooling 1, Experience 2 Interaction	−0.002	3.62
Schooling 2, Experience Interaction	0.0001	0.20
Residence Not in SMSA	−0.145	2.74
Language Variables:		
Spanish Proficiency Indicator	0.056	1.43
English Proficiency[c]		
Direct Effect	−0.307	1.93
Schooling Interaction	0.031	2.54
Experience Interaction	0.012	3.96
Not SMSA Interaction	0.160	2.30
Distance Parameters Between English Proficiency Groups		
Group 1 (origin)	0.000	–
Group 2	0.478	5.13
Group 3	0.689	9.37
Group 4 (scale)	1.000	–
$MSE = 0.240$; Number of Observations = 2576		

[a] The dependent variable is the natural logarithm of earnings divided by weeks worked.

[b] Constant refers to Pacific Division. The full regression has indicators for other divisions.

[c] English proficiency determined by answers to understand and ask questions (Not Asked, Very Well, Well, and Not Well).

NLIN) were used and estimates are given in Table 2. I have omitted the special effects, because once *EP* and *SP* are taken into account, I cannot reject the hypothesis that effects of the special variables are jointly zero. This is a confirmation of the findings of myself with Gould and Welch, using a more cleanly defined language variable. The estimates of the common effects are very similar to the baseline regressions reported in both my dissertation and my article with Gould and Welch.

The effect of *SP* is positive, as expected, but only significant at about the 15 percent level. Holding all other personal characteristics constant, the knowledge of Spanish increases earnings by 5.6 percent for Hispanic men. This indicator really only distinguishes between men in the highest *EP* group who had a Spanish childhood and those in that group who had an English childhood. The finding suggests that, among Hispanic men who are maximally fluent in English, men with a Spanish childhood earn 5.6 percent higher wages. This result is merely suggestive, however, since these "bilinguals" may simply be more adept at integrating into the U.S. labor market.

The effects of *EP* are also as expected. Maximal proficiency increases the return to a year of schooling by 3.1 percentage points and the return to an additional year of experience by 1.2 percentage points. The direct effect of *EP* is negative, but when viewed alongside the positive schooling, experience, and not SMSA interactions, the net effect is positive throughout the range of the data.

The bottom portion of Table 2 presents estimates of the distance parameters between the four *EP* groups. These can be considered

TABLE 3—AVERAGE WAGE, SCHOOLING, POTENTIAL EXPERIENCE, AND SMSA RESIDENCE BY *EP* GROUP FOR HISPANIC MEN, 1975

	English Proficiency (*EP*) Group			
Variable	Not Asked	Very Well	Well	Not Well
English Proficiency[a]	1.000	.689	.478	.000
Weekly Wage	\$221.89	\$199.28	\$183.64	\$151.22
Years of Schooling	11.9	11.2	9.0	7.2
Potential Experience	15.6	18.0	23.6	24.7
Percent Not SMSA	45	45	42	29
Number	700	1118	432	326

[a]Estimated from wage equation.

as estimates of the average levels of *EP* in the cells. The index origin is set by group 1 ($EP = 0$), and the scale by group 4 ($EP = 1$). If group 1 is totally lacking in English proficiency and group 4 is fluent, then group 2 is 47.8 percent as fluent as group 4, and group 3 is 68.9 percent as fluent as group 4 on average.

C. *The Cost of Language Disparity*

In my article with Gould and Welch and in my dissertation, it was shown that Hispanic men in *EP* group 4 have earnings insignificantly different from Anglos with the same schooling and potential work experience, and resident in the same geographic region. Any cost in terms of production foregone must thus be associated with the men in the other three *EP* groups. To measure the cost of language disparity, I first estimate the reduction in marginal product experienced by Hispanic men in these three groups.

Table 3 gives wage, schooling, experience, and SMSA residence comparisons for Hispanic men with English language deficiencies. The first row of Table 3 gives the *EP* estimated in the wage equation of Table 2, the distance parameters between groups. The difference in wage between the Very Well and the Not Well *EP* groups is \$48.07, or 32 percent of the Not Well wage. The weekly wage of the Very Well group is 1.32 times the wage of the Not Well group. In addition, the wage of the Well group is 1.21 times the wage of the Not Well group.

An interesting comparison is in the ordering of the three groups with respect to years of schooling and years of potential work experience. The Very Well group is the most schooled, with 11.2 years, and the least experienced, with 18.0. The Not Well group is the least schooled, 7.2 years, and the most experienced, 24.7 years. The Well group is in-between in both schooling, 9.0 years, and experience, 23.6 years. The greater the English proficiency, the younger and more educated the group on average. Finally, while 45 percent of the Very Well group and 42 percent of the Well group are not SMSA residents, only 29 percent of the Not Well group are.

In Table 4, I present my calculation of the potential gain in production from raising the *EP* of the three deficient groups all to 1.0. The first two rows of Table 4 present the estimated *EP* from the distance parameters between groups in the wage regression of Table 2, and the gain in *EP* represented by an increase in English skills to perfect fluency. This gain ranges from .311 for the Very Well group to 1.000 for the Not Well group. The next row gives the *Delta* multiplier of equation (3) evaluated at the means of education, experience, and not SMSA residence for each group. This multiplier is largest for the Very Well group, .328, and smallest for the Not Well group, .260.

The fractional gain in wage from an improvement of English skills to full fluency is computed by multiplying the change in *EP* needed by the *Delta* multiplier. The next row of Table 4 presents the fractional gain in wages of the three groups. It ranges from .102 for the Very Well group to .260 for the Not Well group. This means that moving the

TABLE 4—CALCULATING THE POTENTIAL GAIN FROM ACQUIRING FLUENCY IN ENGLISH BY *EP* GROUP: HISPANIC MEN, 1975

	English Proficiency (*EP*) Group		
Variable	Very Well	Well	Not Well
English Proficiency[a]	.689	.478	.000
Gain in English Proficiency to Fluency	.311	.522	1.000
Delta Multiplier[b]	.328	.323	.260
Effect as Percent of Wage of Fluency	.102	.169	.260
Dollar Value per Week	$20.23	$31.04	$39.32
Value per Year (50 weeks)	$1016.50	$1552.00	$1966.00
Years to Retirement	29	26	27
Present Value of Fluency[c]	$19,505.06	$27,744.86	$36,030.94
Number of Men	1118	432	326
Aggregate Value	$21,806,657	$11,985,780	$11,746,086

[a]Estimated from wage equation.
[b]Estimated from wage equation assuming average variable values.
[c]At 3 percent.

average Very Well group member up to full English fluency will raise his wages by about 10 percent, moving the average Well member to fluency will raise his wages by about 17 percent, and moving the average Not Well member's raises his wages by about 26 percent.

Starting from the average weekly wages in Table 3, the next row of Table 4 presents the dollar value per week of the move to fluency. This value per week ranges from $20.33 for the Very Well group to $39.32 for the Not Well group. Assuming a 50-week working year, I next compute the dollar value per year of English fluency by group. This ranges from about $1000 for Very Well to nearly $2000 for Not Well. To compute a lifetime present value of the proficiency improvement, I estimate the average years of working life for each group based on their age and retirement at 65-years of age. I assume that the dollar value per year continues each year until retirement, and I discount future values at a 3 percent annual rate to get a net present value amount.

The upshot is that the present value of acquiring fluency is $19,505.06 for Very Well, $27,744.86 for Well, and $36,030.94 for Not Well. These are my per person estimates of the cost of language disparity. The aggregate cost associated with the Hispanic men in my sample is computed by multiplying the per person cost by the number of men in each group and summing across groups. I estimate the aggregate cost of language disparity associated with the Hispanic men in my sample at over $45 million. Since Spanish speakers account for a little less than one-half of all English deficient persons, this $45 million figure probably understates the full cost to society of language disparity.

III. Summary and Implications

I have explored theoretical models of the frictions caused by language disparity and estimated the cost of language disparity for Hispanic men in the United States. Two basic types of theoretical models were outlined, based on the nature of the product. For personal products, the identity and language of the producer (or seller) becomes important to the consumer. Differences in price (and income) between language groups can emerge due to differences in market conditions, as can premiums for bilinguality and knowledge of modal languages by minority language group members.

For impersonal products, the identity of the producer is irrelevant to the consumer, so the motivation for language effects comes on the production side. Language of production is important only if there is an incentive for integration of disparate language groups, and

only if integration is enhanced by workers being able to communicate with each other.

In the empirical application, I estimated the cost of English deficiency for Hispanic men in the United States. I estimated the present value of their lifetime cost of not being fluent in English as ranging from \$19,000 for the most nearly fluent to \$36,000 for the least.[7] This means that, if it costs less than \$19,000 to give an Hispanic nearly fluent in English additional English skills through training to become fluent, then there could be a net gain to society. Similarly, if the cost of training Hispanics from the least nearly fluent in English group is less than \$36,000 per person, then there is a potential net gain from doing so. Less training than necessary to raise Hispanic men to full English fluency might also provide net gains if the gain exceeds the costs. The maximum gains I estimated give upper limits on the effectiveness of language training.

If these personal gains from acquiring English fluency are so substantial, why have not the Hispanics themselves done so? The simplest explanation is that their costs of acquiring English fluency are greater than their benefits. This might be especially important for older immigrants. Another reason could be the existence of a borrowing constraint making it difficult for Hispanic men to finance English language training. Also, the gains were calculated using a 3 percent real discount rate, but the personal discount of Hispanics could be much higher. Illegal or temporary immigrants especially would discount future returns to English skills.

The decision of whether the estimates of cost reported here are convincing is left to the reader. I have built on work begun by myself and others that suggests that language differences go a long way toward explaining differences in economic success. The estimates of costs of language disparity reported here are not the last word, but researchers in this area will at least have to address the results described here.

My estimates show that, once *EP* and *SP* are taken into account, the differences in wages associated with ethnicity, nativity, schooling abroad, and time in the United States are no longer statistically significant. I do not view this as evidence that these factors are unimportant, rather that their effects are mediated through language characteristics. Indeed, in earlier work I explored the determinants of *EP*. The U.S.-born Hispanics show greater proficiency than immigrants, and proficiency increases with schooling. Among immigrants, those arriving as preschool children are most proficient, immigrants arriving after beginning but before completing schooling are next, and those arriving after completing schooling are least proficient. Immigrants increase proficiency with schooling and experience in the United States.

The bilingual premiums I estimated are transitional phenomena—if all Hispanics are made fluent in English, then bilinguals will suffer earnings losses. These losses are not taken into account, and probably ought not to be.

A final qualification of the work reported here concerns language minority scale effects. If Hispanics become perfect substitutes for Anglos, then the wages of laborers well tend to fall somewhat, depending on how large the Hispanic labor force is relative to the Anglo. This makes the estimates of cost based on constant marginal products of the two groups an overstatement. In addition, the models developed suggest that opportunities for production clustering might permit agglomeration economies for those who do not speak modal languages. My attempts at empirically modeling such economies have been inconclusive, and my estimates do not take account of such economies.

[7] I assume this represents differences in true productivity. To the extent that the language-related wage differences between Hispanics reflects discrimination rather than true productivity differences, the net gain to society will be overstated.

REFERENCES

Grenier, Gilles, "The Effect of Language Characteristics on the Wages of Hispanic-American Males," *Journal of Human Resources*, Winter 1984, *19*, 25–52.

McManus, Walter, "Effects of Language Char-

acteristics on Earnings: Hispanic Men in the United States," unpublished doctoral dissertation, University of California, Los Angeles, 1983.

______, **Gould, William and Welch, Finis,** "Earnings of Hispanic Men: The Role of Proficiency in the English Language," *Journal of Labor Economics*, April 1983, *1*, 110–30.

Reimers, Cordelia W., "Labor Market Discrimination Against Hispanic and Black Men," *Review of Economics and Statistics*, November 1983, *65*, 570–79.

U.S. Department of Commerce, Bureau of the Census, *1980 Census of Population*, Vol. 1, Part 1, Washington: USGPO, 1983, ch. C.

______, *Survey of Income and Education*, Microdata file (computer tape), 1976.

Equilibrium Relationships Between Money and Other Economic Variables

By JAMES R. LOTHIAN*

Central to the quantity theory of money are a number of important propositions about the long-run equilibrium effects of changes in the nominal stock of money on other economic variables. A standard way of testing these propositions is to use quarterly or annual data, explicitly model the lag structure and then derive the long-run solution of the model from the empirical estimates. An alternative is to use some type of smoothing procedure to approximate positions of equilibrium and then to use these transformed data directly in testing hypotheses. The National Bureau technique of averaging data over reference cycle phases is one such method; Robert Lucas's application of Fourier transforms in "Two Illustrations of the Quantity Theory of Money" (1980) is another; and John Geweke's method (1982) of frequency decomposition is a third.

An entirely different way of approaching the problem is to use cross-country-average rather than time-series data as the basic units of observation. The advantage, according to Lucas, is that, "Since the two quantity-theoretic laws [that he examines] are obtained as characteristics of steady states, or limiting distributions, of theoretical models, the ideal experiment for testing them would be a comparison of long-term average behavior across economies with different monetary policies but similar in other respects" (p. 1006).

In this paper I conduct such an experiment. The data that I use are for 20 OECD countries over the period 1956–80. The specific relationships that I examine are those between money and the price level, money and real income, money and interest rates, and money and exchange rates. In the main the data accord well with the quantity-theoretic model. Classical neutrality holds. There is evidence of a Fisher effect, albeit a less than complete effect, on interest rates. Finally, the data are consistent with long-run purchasing power parity and, hence, correspondingly with a long-run monetary approach to exchange rate determination.

I. Theory and Data

The theory underlying the empirical investigation summarized in the next section of the paper is the quantity theory of money and its corollary in a multicountry context, the monetary approach to exchange rate determination. For an extended treatment of the basic quantity theory, the interested reader can refer to Milton Friedman (1969); for a statement of the monetary approach to exchange rates, to Jacob Frenkel (1976). Given the purpose of this paper, I merely illustrate the implications of the two by considering a simple example in which I assume an initial steady-state equilibrium that is interrupted by an unanticipated one-time increase in the rate of growth of the domestic money supply. I then trace the short-term and long-term implications of that change for other domestic economic variables and for the exchange rate.

To proceed with the example, suppose that the money supply had been increasing for an extended period of time at a rate of 5 percent per year and that it then undergoes a sudden, one-time acceleration to a rate of growth of 10 percent per year. Suppose further that when money had been growing at that 5 percent rate, inflation had averaged 2 percent, real growth 3 percent and the yield on long-term government bonds 5 percent.

The initial effect is to create a disparity between the rate at which individuals wish to

*Vice President, Citicorp Investment Bank, 55 Water St., New York, NY 10043. I am grateful to Michael Darby, Cornelia McCarthy, and Frederick Sturm for comments, and to Donna Bettini and Barbara Podesta for their assistance.

accumulate money balances and the rate at which they are actually doing so. The amount of money individuals actually hold relative to their nominal incomes, therefore, becomes greater than the amount they desire to hold. Correspondingly, the implicit yield on those money balances on the margin falls below the yields on bonds and on other alternative assets. To correct these imbalances, individuals increase their spending. In the process, yields on other assets fall, output grows more rapidly, and the rate of inflation gradually begins to increase.

Within the context of the quantity theory, the downward effect on rates of interest and upward effect on the growth of real output are temporary phenomena. In the final equilibrium position, the rate of inflation will have risen to 7 percent, increasing by the same amount as the rate of monetary growth, and bond yields will have risen to 10 percent, reflecting the now higher actual (and anticipated) rates of inflation. The rate of growth of real output will be unaffected.[1]

With regard to exchange rates there is an immediate analogue. The initial effect of a sudden and unanticipated monetary acceleration in lowering domestic real rates of interest will, everything being equal, lead to an incipient capital outflow. To offset that outflow, the exchange rate must depreciate enough to produce anticipation of a subsequent appreciation. That, in turn, will lessen, or eliminate, the potentially widened differential between the yields on domestic and foreign assets. In the final equilibrium position, overshooting of this sort will be eliminated and the rate at which the exchange rate is changing will bear a one-to-one relationship to the change in the rate of growth of the domestic money supply.

So, to build on the example given above, again assume that the rate of growth of the domestic money supply increased by five percentage points. Assume further that the rate of growth of foreign-country money supply remained unchanged. In the new equilibrium, the exchange rate, the domestic currency price of a unit of foreign-country currency, would now increase at a five percentage points per annum faster pace.[2]

These, of course, are highly simplified examples. For my purposes, however, simplicity of this sort is a virtue. If the data do actually support these simple formulations, one can have greater confidence in the underlying theory as a useful tool.

To test the various hypotheses just outlined, I assembled annual data for 20 OECD countries for the period 1956–80 for an *M*1-type definition of money, consumer prices, real income (*GNP* or *GDP*, depending upon availability), U.S. dollar exchange rates, and, for a 14-country subset, long-term bond yields. The source of most of these data was the International Monetary Funds's (IMF) *International Financial Statistics* and companion tapes. The remainder came from the NBER data base described in the appendix to Michael Darby, myself et al. (1983).

To minimize the effects of shorter-term fluctuations, I computed the average rates of growth of each variable for each country for the two subperiods 1956–73 and 1974–80 (see Table 1). I then computed the changes in these average rates of growth from the one period to the next and, with the one exception noted below, used them as my basic units of observation. For *M*1, real income, the price level, and the exchange rate, these were shifts in average annual logarithmic rates of change; for the interest rate variables used in the examination of exchange rate relationships, they were shifts in average annual arithmetic rates of change. In each of these instances, therefore, the end result was a measure of longer-term acceleration, a "growth shift." The exception to this rule was the interest rate variable used in the Fisher equation, which for theoretical rea-

[1]The transition to equilibrium following these initial effects is harder to specify. Overshooting of inflation, a one-time shift in the desired ratio of money to income, and a decline in the rate of growth of real income are all features of the process.

[2]Implicity assumed is no permanent alteration in the growth rates of the quanties of real cash balances demanded in the two countries. A one-time decline in the level of real cash balances demanded in the domestic country, and hence, temporarily lower growth rate, will characterize the transition process period.

TABLE 1—AVERAGE ANNUAL RATES OF CHANGE AND STANDARD DEVIATIONS: 20 OECD COUNTRIES, 1956–80

	Means[b]		Standard Deviations[b]	
Variable[a]	1956–73	1974–80	1956–73	1974–80
*M*1	9.11	11.51	6.51	8.52
Consumer Price Index	4.30	11.68	3.06	8.62
Real Income	5.10	2.61	2.54	2.63

Sources: IMF and Darby, myself et al.
[a]Percentage rates of change were computed as first differences of the natural logarithms of the variables multipled by 100.
[b]Shown in percent.

sons I constructed as the between-period shift in the average *level* of interest rates.[3]

The first subperiod is the era of Bretton Woods, 1973 marking the point of its final demise. The second is the era of managed floating. In addition to facilitating the study of exchange rate movements per se, this division offers a rather unique opportunity of another sort. After the breakdown of the Bretton woods system, the monetary policies of the various countries included in this study became more divergent. The result is, therefore, a much richer body of data from which to make inferences than the period of fixed exchange rates alone affords.

II. Empirical Evidence

In discussing the empirical findings, I divide the four relationships into three categories, lumping the money-price and money-real output relationships under the common heading of "classical neutrality," and then going on to consider the relationship between interest rates and inflation rates, the Fisher equation. Exchange rate relationships make up the final subsection. Throughout, the emphasis is on graphical evidence in which the theoretical and the hypothesized relationships are compared. Related regression results and formal statistical tests of hypotheses are presented in Table 2.

A. *Classical Neutrality*

Some preliminary evidence with respect to neutrality is contained in Table 1. The first two columns in the table present the averages of the annual rates of growth of real income, consumer prices, and *M*1 for the 20 countries for the separate subperiods 1956–73 and 1974–80. The second two columns show the corresponding standard deviations.

Consider the averages first. For the 20 countries taken together, we see an increase in both monetary growth and inflation from one period to the next, but a decrease in the rate of real income growth. That's one small bit of evidence in favor of the quantity-theoretic positive correlation between monetary growth and inflation. But, except for contradicting the naive Phillips curve, it tells us little about neutrality. More important on that score is the pattern of variability between periods.

With the final breakdown of the Bretton Woods system of fixed exchange rates in 1973, foreign countries gained greater policy independence. Monetary growth, not surprisingly, became more variable. So too did inflation. The variability of real income growth, in contrast, remained the same. The implications, therefore, are entirely consistent with the neutrality proposition: longer-term changes in monetary growth apparently affect the rate of inflation, but not the rate of real income growth.

[3]Differentiating the standard Fisher equation with respect to time results in an equation linking the increase in the interest rate and the increase in the anticipated rate of inflation. The empirical analogue here is a relationship between the shift in the average level of interest rates and the shift in the average rate of growth of the price level.

TABLE 2—REGRESSIONS FOR CROSS-COUNTRY GROWTH SHIFTS: 20 OECD COUNTRIES, 1956–80

$$Z=\beta_0+\beta_1 X_1+\beta_2 X_2$$

Variables[a] Dependent	Independent (X_1, X_2)	DF No. of Observations	Coefficients[b] β_0	β_1	β_2	$\bar{R}^2$/ SEE	Test Constraint	F-Ratio	Significance Level
P = Cost of Living Index	$M1$ = Narrowly Defined Money	18	.052	.891		.665	$\beta_1=1$	.053	.480
			(.008)	(.149)		.033			
y = Real Income	$M1$	18	−.027	.097		.121	$\beta_1=0$	3.606	.071
			(.003)	(.051)		.011			
i = Long-Term Bond Yield	P	14	.008	.541		.634	$\beta_0=0$;	23.412	0.000
			(.007)	(.112)		.012	$\beta_1=1$		
e = Spot Dollar Exchange Rate	DP	17	−.004	.898		.871	$\beta_0=0$;	1.457	.260
			(.005)	(.084)		.020	$\beta_1=1$		
e	$DM1$; Dy	16	−.048	.914	−1.397	.625	$\beta_0=0$;	2.038	.162
			(.031)	(.178)	(.769)	.035	$\beta_1=1$		
e	$D(M1-y)$	17	−.006	.889		.591	$\beta_0=0$;	.687	.521
			(.008)	(.171)		.171	$\beta_1=1$		
e	$D(M1-y)$; Di	10	.002	.403	.053	.588			
			(.007)	(.186)	(.014)	.022			

Sources: See Table 1

[a] D denotes a variable constructed as the difference between the relevant foreign and U.S. variables. With the exception of interest rates, all variables are differences between the average annual logarithmic changes in the period 1974–80 and 1956–73. The interest rate term in the last regression is the difference between the average annual arithmetic changes in interest rates in the two periods for the foreign country less the similar difference for the United States. The interest rate term in the third regression is the difference in the average levels of interest rates in the two periods.

[b] Standard errors are shown in parentheses.

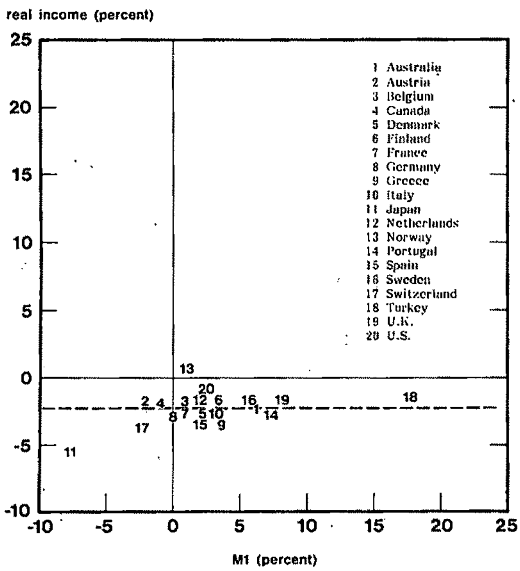


FIGURE 1. GROWTH SHIFTS: REAL INCOME VS. $M1$

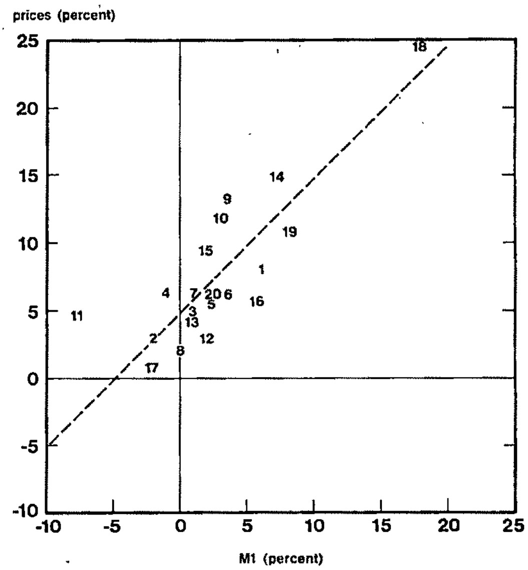


FIGURE 2. GROWTH SHIFTS: PRICES VS. $M1$

These relationships, moreover, are not just a statistical quirk caused by combining the data for all 20 countries. This becomes apparent when Figures 1 and 2 and the corresponding regressions are considered. Figure 1 plots the growth shift in money in each country, the change from one period to the next in the average annual rate of monetary

growth, against the similar growth shift in real income. Figure 2 plots the growth shift in money against the growth shift in the price level.

The money-real income points are scattered about a horizontal line drawn through the means of the observations. All countries experienced roughly the same decrease in real growth despite huge differences in the change in the average rate of monetary growth. The money-inflation points, however, are scattered about a 45° line, again drawn through the means. If monetary growth in one country accelerated by, say, five percentage points more than in another, there was a similar difference in the amount by which inflation accelerated.

The regressions add to the story told by the two figures. Growth shifts in real income are statistically unrelated to those in $M1$: we cannot reject the quantity theory hypothesis that the intercept and slope coefficients jointly are zero. Growth shifts in prices and $M1$, however, are closely matched: we cannot reject the hypothesis that the slope coefficient in this equation is unity. The data, therefore, clearly support the classical neutrality proposition of the quantity theory.[4]

B. *The Fisher Equation*

Figure 3 plots the shift in the average level of the long-term bond rate against the growth shift in prices (the shift in the average rate of inflation) for the 14 out of 20 countries for which there are interest rate data. If the Fisher relation held exactly and the average rate of inflation was an accurate proxy for the anticipated rate, the points would all fall on a 45° line through the origin. A permanent increase in the anticipated rate of inflation of, say, five percentage points would, as in the example outlined in the previous section, increase the nominal bond yield by five percentage points.[5]

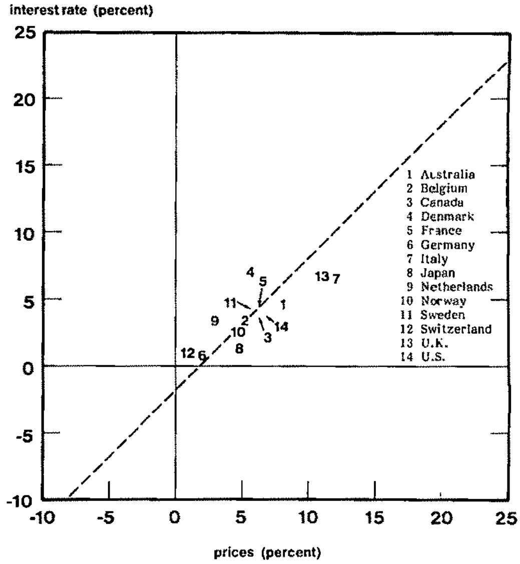


FIGURE 3. GROWTH SHIFTS: INTEREST RATE LEVEL VS. PRICE GROWTH

In fact, most of the points do cluster about a 45° line, but the relationship is far from perfect and the line itself does not intersect the vertical axis at the origin. Hence, even relative to the substantial average increases in both rates of inflation and rates of interest (over five and roughly four percentage points, respectively), there is a considerable dispersion of the observations about the line. This, in turn, suggests that real rates of interest, measured as they implicitly are here as the difference between the average nominal bond yield and the average actual rate of inflation, were by no means constant among countries. Nor were they constant over time within countries. In most of the countries in the sample, the increase in inflation exceeded the increase in bond yields by one percentage point or more. In the estimated regression equation, this variability in real rates over both time and space leads to a rejection of the joint hypothesis of a zero intercept and a unit slope coefficient.[6]

[4] There is the possibility of a more complex relationship. Friedman (1977) has associated the decline in real income growth in many of these countries with an increase in the variability of inflation and ultimately, therefore, in monetary policy.

[5] The discussion ignores both the positive tax effect on interest rates described by Darby (1975) and the negative Mundell effect.

[6] See Frederic Mishkin (1982, 1984) for evidence drawn from multicountry data on the nonconstancy of real rates across countries and over time within countries, respectively.

One factor that may be important here is a decline in the *ex ante* real rate. The average rate of real growth declined in almost all of the countries between the first period and the second period. Using the real rate of growth as a proxy for the real return on real assets (Friedman and Anna Schwartz, 1982, ch. 10), we might, therefore, infer that the *ex ante* real yield on bonds has also declined.

That, however, cannot be the full explanation since there is little correlation among countries between the degree of movement in our measure of *ex post* real bond yields and real rates of growth. An alternative explanation is that in most of the countries in the sample, there was simply a lag in the adjustment of expectations to changes in rates of inflation. On the surface, this hypothesis seems implausible: the time periods involved appear too long for rational market participants not to have adjusted their expectations fully. The objection to that line of reasoning, though, is that it implies a knowledge of the inflation process on the part of individual market participants that they may not have had. Predicting longer-term rates of inflation is essentially a problem of predicting the longer-term rate of monetary growth. The latter, in turn, is a question of the monetary policy regime. Given inertia in the political process, formation of expectations in a regressive and seemingly myopic fashion may, therefore, be entirely rational. In the absence of an obvious and dramatic political change, market participants may view policy over the next decade as differing little from policy in the last.[7]

To sum up, I find support for the Fisher equation as a general proposition. Conversely, the data provide no support for the popular notion that easy money and low interest rates go hand in hand. Real rates of interest, as measured here, were far from constant either temporally or geographically, however.

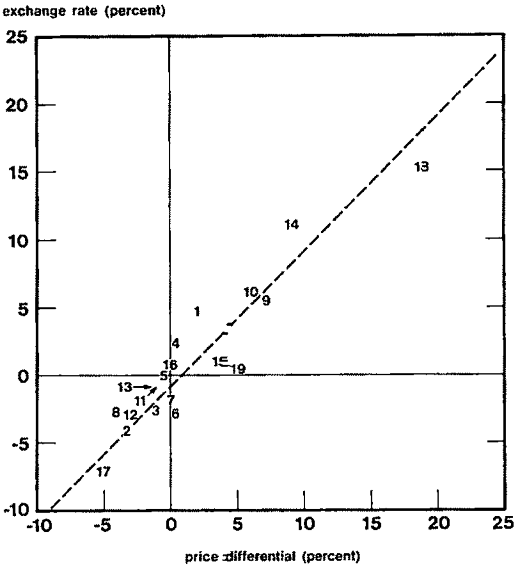


FIGURE 4. GROWTH SHIFTS: EXCHANGE RATE VS. PRICE DIFFERENTIAL

C. *The Monetary Approach to Exchange Rates*

Figure 4 plots the shift in the rate of change of the foreign currency versus dollar exchange rate against the shift in the inflation differential, where the latter is defined as the shift in the difference between the average rates of foreign and U.S. inflation. Figure 5 plots a similar relationship except that the shift in the differential excess rate of monetary growth replaces the shift in the inflation differential as a variable. It, in turn, is defined as the shift in the difference between the average rates of foreign and U.S. $M1$ growth less the shift in the difference between the corresponding rates of real income growth. The first component of this latter algebraic total indicates the degree to which the differential growth rates of the money supplies in the foreign country and the United States has changed; the second proxies the extent to which the differential growth rates of the real quantities of money

[7]See Friedman and Schwartz (p. 556) for a similar discussion of bond yields in the United States in the latter part of the nineteenth century. They attribute the substantial difference between bond yields and the (negative) rate of inflation at that time to investors' fears that agitation for free silver would lead to an abandonment of gold and renewal of inflation. This example, coupled with the results obtained here as well as with recent U.K. and U.S. experience, raises the dual questions of what can be considered rational behavior a priori and how one can ascertain empirically whether behavior is or is not rational.

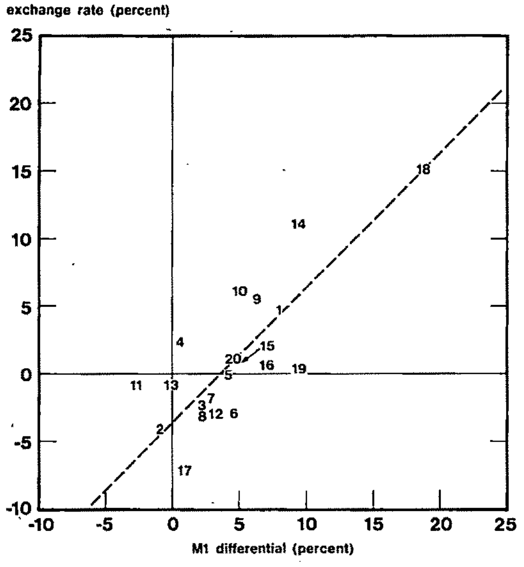


FIGURE 5. GROWTH SHIFTS: EXCHANGE RATE VS. $M1$ DIFFERENTIAL

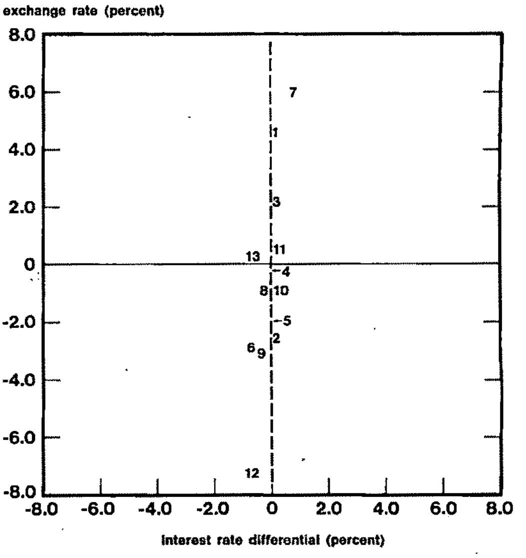


FIGURE 6. GROWTH SHIFTS: EXCHANGE RATE VS. INTEREST RATE DIFFERENTIAL

demanded in the two countries has changed. Combined, the two therefore provide a measure of the shift in the relative excess supplies of monies (or demands if the total is negative).[8]

Figure 4 is, of course, a direct transformation of Figure 5, provided that the money-to-inflation link holds. It is just the purchasing-power-parity relationship stated in terms of accelerations in prices rather than in levels of prices.

As the figures show, both longer-term relationships hold tolerably well. In each instance, the points are clustered about 45° lines drawn through the means of the observations, indicating the existence of one-to-one relationships between shifts in the rates of change of exchange rates and shifts in the rates of change of the other variables.

The regressions corresponding to the charts buttress this conclusion. In all instances, the percentage of variation of the exchange rate variable accounted for is substantial (.87 with the inflation differential; .63 and .59 with the unconstrained and constrained differential monetary growth variables, respectively). In all instances, however, I am unable to reject the joint hypothesis that the intercept term in the regression is zero and the slope term is unity.

An additional implication of the monetary approach is also confirmed by the data. That model posits a positive relation between the exchange rate, defined as it is here as the foreign-currency price of a dollar, and the difference between the foreign and the U.S. interest rate. The rationale is that an increase in the foreign interest rate reduces the real amount of foreign money that foreigners want to hold, and hence for a given money supply produces an excess supply of money. This, in turn, results in upward pressure on the exchange rate or weakening of the foreign currency relative to the dollar.

The common view to the contrary is that an increase in a country's interest rate neces-

[8]It is only a proxy since other arguments of the demand for money functions are omitted and the income elasticities of demand implicitly are assumed to be unity. The regression results reported in Table 2 are broadly consistent with this second assumption. The income elasticity estimates derived from the sixth equation are not significantly different from unity. Additional results for the smaller sample indicate, however, that interest rates also clearly matter. The interest rate term in the seventh regression has the correct sign and is significantly different from zero.

sarily strengthens its currency via its effects on capital inflows. The data do not support this second proposition as a long-term hypothesis. As Figure 6 indicates, there is no relation between the growth shift in the exchange rate and the shift in the differential rate of change of nominal bond yields. Conversely, when the interest rate variable is included in a regression equation representative of the monetary approach, it is statistically significant and has the expected positive sign. An increase in the own-country interest rate, other things being equal, weakens rather than strengthens its currency.

III. Conclusions

In this paper I have examined three sets of hypotheses associated with the quantity theory of money: the classical neutrality proposition, the monetary approach to exchange rates, and the Fisher equation. The data are completely consistent with the first two and moderately supportive of the last. Given the nature of the data, the fact that they are generated by, to use Lucas's term, an "ideal experiment," and given the simplicity of the tests, I view the results as a strong confirmation of the theory.[9] Looked at from the opposite perspective, they provide little or no support to the corresponding alternative hypotheses, the naive Phillips curve, the interest rate theories of exchange rate determination and the liquidity preference theory of interest rate determination.

[9]For corroborative cross-country evidence for the money-price relationship see Schwartz (1973) for a sample of 40 developed and less developed countries, and William Norton and Robin McDonald (1981) for a sample of 10 developed countries. Lucas, as an introduction to his U.S. analysis, presents similar evidence for 16 Latin American countries.

REFERENCES

Darby, Michael R., "The Financial and Tax Effects of Monetary Policy," *Economic Inquiry*, June 1975, *13*, 266–76.

______, **Lothian, James R. et al.**, *The International Transmission of Inflation*, Chicago: University of Chicago Press, 1983.

Frenkel, Jacob A., "A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence," *Scandinavian Journal of Economics*, April 1976, *78*, 200–44.

Friedman, Milton, "The Optimum Quantity of Money," in his *The Optimum Quantity of Money and Other Essays*, Chicago: Aldine, 1969.

______, "Nobel Lecture: Inflation and Unemployment," *Journal of Political Economy*, June 1977, *85*, 451–72.

______ **and Schwartz, Anna J.**, *Monetary Trends in the United States and the United Kingdom*, Chicago: University of Chicago Press, 1982.

Geweke, John, "Measurement of Linear Dependence and Feedback Between Multiple Time Series," *Journal of the American Statistical Association*, June 1982, *77*, 304–13.

Lucas, Robert E., Jr., "Two Illustrations of the Quantity Theory of Money," *American Economic Review*, December 1980, *70*, 1005–14.

Mishkin, Frederic S., "Are Real Interest Rates Equal Across Countries? An Empirical Investigation of International Parity Conditions," Working Paper No. 1048, National Bureau of Economic Research, December 1982.

______, "The Real Interest Rate: a Multi-Country Empirical Study," *Canadian Journal of Economics*, May 1984, *27*, 283–311.

Mundell, Robert A., "Inflation and Real Interest," *Journal of Political Economy*, June 1963, *71*, 280–83.

Norton, W. E. and Robin McDonald, "Implications for Australia of Cross-Country Comparisons in Economic Performance," *Economic Record*, December 1981, *57*, 301–18.

Schwartz, Anna J., "Secular Price Change in Historical Perspective," *Journal of Money, Credit and Banking*, February 1973, *5*, 243–69.

International Monetary Fund, *International Financial Statistics*, various issues.

The Market for "Lemons" Reconsidered: A Model of the Used Car Market with Asymmetric Information

By JAE-CHEOL KIM*

The purpose of this paper is to analyze the working of the used car market, a market suffering severe informational asymmetry, by generalizing in two respects works of George Akerlof (1970) and Charles Wilson (1980) in the Walrasian paradigm.[1]

First, although Akerlof correctly points out a possible market failure in the used car market, his description of it overlooks the ability of each agent to freely choose whether to be a buyer or a seller. In other words, in the used car market, unlike other markets with informational asymmetry—the insurance market and the labor market—an agent can change his position from buyer to seller, or vice versa, with little or no transaction cost. Instead, Akerlof (and Wilson) arbitrarily divides the agents in the market into two groups, buyers and sellers, where, for example, an agent in the buyer group is supposed to buy only a used car. This is unnecessarily restrictive. It is possible that, if the price of used cars goes up, a used car buyer may want to shift his demand from a used car to a new car, and as a result he will be a used car seller.

Second, my model treats the quality of used cars as an endogenous variable in contrast with the Akerlof-Wilson model where each seller is exogenously endowed with a car of given quality. However, it is well recognized that the quality of a car depends not only on purely stochastic elements as perceived by Akerlof and Wilson, but also on its owner through endogenous factors such as maintenance, driving habits, and the like. This paper places more emphasis on the latter by assuming that the quality of a car is a function of the maintenance level.

Throughout the analysis, however, I will preserve the informational structure of Akerlof and Wilson by assuming that no activities of information acquisition and transmission (i.e., signaling, warranty, or search) are allowed. As a result, potential buyers make their decision based only on the average quality of used cars.

The basic model is constructed in Section I as follows. A car lasts two periods; it is new in the first and used in the second. There are many types of agents in the market and each agent can buy either a new car or a used car depending on his (or her) preference structure. The buyer may also refrain from buying a car. If he buys a new car, he chooses a level of maintenance for it, which determines a service flow. The service flow of a used car (the quality of a used car) depends on maintenance in the previous period when the car was new, but not on that in the current period. An owner of a new car can either sell or keep it after one period when it becomes used.

In Section II, given this setting, I characterize an equilibrium in the used car market, examining two contrasting cases depending on the underlying structure. An interesting result obtained in the first case is that the Akerlof's Lemons Principle need not hold in the used car market. Akerlof originally asserts that the lowest quality car will drive out cars of higher qualities. Yet the situation is not that extreme. Even in Akerlof's example, a nontrivial equilibrium could be shown to exist by slightly changing the assumed parameters. (Wilson shows this in a more general framework.) Then, the principle can be reinterpreted as saying that the

*Department of Economics, Amherst Campus, State University of New York, Buffalo, NY 14260. I thank many for their helpful comments, but am especially indebted to James Friedman, David Gordon, and Dave Weimer, and the anonymous referees who made valuable suggestions on an earlier version of this paper.

[1] Wilson provides two more conventions used to set the price in addition to the Walrasian convention for the analysis of market equilibrium with asymmetric information.

average quality of nontraded cars is higher than that of traded cars. Recently, Eric Bond (1982) tests this hypothesis for the pickup truck market and rejects it. His finding is that there is no significant difference in quality between traded and nontraded trucks. This paper provides a theoretical support to this finding by showing that either quality can be higher than the other. However, this is not because there is a mechanism to reduce or eliminate asymmetry in information as Bond himself conjectures. Rather, this is because there may be some people who value car service so highly that they maintain their cars exceptionally well when they are new in order to keep the service flow at a very high level. This leaves good used cars after one period. However, they will sell their good used cars and buy new cars in order to continue receiving the very high service flow associated with new cars. Also in Section II, the second case is considered. Although the case appears to be very similar to the Akerlof-Wilson model in many respects, a significant discrepancy is shown to exist. The concluding section summarizes the results and suggests areas for further research.

I. The Basic Model

Consider the used car market with many types of agents and asymmetric information. A buyer of a used car can observe only the average quality of used cars while a seller knows the quality of his own car. Let x be an index denoting the quality of a car, measuring its overall efficiency in terms of factors such as driving performance, the number of breakdowns including the resulting inconvenience and cost of repair, comfort, and so on. A unique feature of the model is that the quality of a car is endogenous, varying with a level of maintenance, m, which represents the expenditure on a car for any preventive purpose such as regular checkups.

When a car is new, its quality, $x_n(m)$, is a continuously differentiable, increasing and strictly concave function of the maintenance level; $x_n'(m)>0$ and $x_n''(m)<0$. However, it is assumed that maintenance on a used car has no effect on quality. Rather, the quality of a used car is solely determined by the maintenance level in the previous period when the car was new. Let $x_u(m)$ be the quality of a used car where m is the maintenance level applied in the previous period. Also assume that $x_u'(m)>0$ and $x_u''(m)<0$. Finally, for analytic simplicity, it is assumed that $x_u(\infty)<x_n(0)$, that is, any used car is lower in quality than any new car. Each agent is characterized by a real number t distributed on $T=[\underline{t},\bar{t}]$ with a density function $w(t)>0$ where $\underline{t}>0$. A type-t agent tries to maximize two-period expected utility and has the following von Neumann-Morgenstern one-period utility function, being risk neutral with respect to quality:

$$U(x,e:t)=tx-e, \tag{1}$$

where e is the expenditure on a car.[2] The ownership of no car is equivalent to owning a car of zero quality. If an agent buys a used car, his expected utility is a linear function of the average quality of used cars offered in the market. Finally, let P_n, P_u and β be the price of new cars, the price of used cars and a discount parameter, respectively.

Now let us consider the optimum behavior of a type-t agent. Assume that at the beginning of the first period, he has no car. Given this state of "having no car," he faces four possible choices: 1) buy a new car, maintain it and in the second period sell it; 2) buy a new car, maintain and keep it for both periods; 3) buy a used car and sell it in the second period; and 4) do not buy a car. It is simple to calculate the two-period utility, $S_2(t)$, if option 2) is taken,

$$S_2(t)=-P_n+Z_2(t), \tag{2}$$

where $Z_2(t)=tx_n(m_2(t))+\beta tx_u(m_2(t))-m_2(t)$ is the two-period net service from a

[2] The utility function in a general context can be defined as $U(x,c:t)=tx+c$, where c is consumption other than car service. Then, letting y be per period income, $c=y-e$. It can be easily seen that income level has no effect on optimal decision in this simple model if each agent receives the same rate of income, irrespective of agent type. Thus, without losing generality, we may set y equal to zero, which gives the utility function in equation (1).

car when adopting option 2 for which $m_2(t)$ is the optimum maintenance. If he chooses options 1, 3, or 4, he goes back to the starting state of having no car after one period. Therefore, if an option is optimal for him in the second period, so must it be in the first period. This means that we only have to consider stationary policies taking the same option in each period.

Let $Z_1(t) = (1+\beta)(tx_n(m_1(t)) - m_1(t))$ and $Z_3(t) = (1+\beta)tx_u^e$ be the two-period net service when adopting options 1 and 3, respectively, where $m_1(t)$ is the optimum maintenance for option 1 and x_u^e is the expected average quality of used cars in the market. It is assumed that each agent has the same expectations about the average quality. Letting $S_i(t)$ be the two-period utility, a type-t agent can obtain when taking option i, $(i=1,3,4)$,

$$(3) \quad S_1(t) = (-P_n + \beta P_u)(1+\beta) + Z_1(t),$$

$$S_3(t) = -P_u(1+\beta) + Z_3(t),$$

$$S_4(t) = 0.$$

Let $V(t)$ be the maximum two-period utility of a type-t agent. Then, from the above discussion, $V(t)$ is an envelope of the $S_i(t)$'s; that is,

$$(4) \quad V(t) = \max\{S_i(t),\ i=1,2,3,4\}.$$

Now the simple nature of the problem immediately implies the following properties. First, noting that S_i's depend on P_n, P_u, and x_u^e, let us define T_i using the above equation as follows:

$$(5) \quad T_i \equiv T_i(P_n, P_u, x_u^e) = \{t: V(t) = S_i(t)\}, \quad i=1,2,3,4.$$

Then T_i is a set of agent types who maximize utility by adopting option i, $i=1,2,3,4$. There may be an interval of t over which two or more S_i's coincide. For analytic simplicity, but without sacrificing essentials, I will assume away this possibility.[3] Also, if some t belongs to more than one T_i, then I include it in T_i with the smallest index. An agent in T_1 buys a new car in each period and sells it after one period. Therefore, T_1 is a set of used car sellers. Similarly, T_3 is a set of used car buyers. Each agent in T_2 buys a new car and keeps it for both periods so that he is an owner of a nontraded used car. T_4 is a set of agents who do not buy cars.

Second, $m_1(t)$ maximizes $tx_n(m) - m$ while $m_2(t)$ maximizes $tx_n(m) + \beta tx_u(m) - m$. Assuming interior solutions, $m_1(t)$ and $m_2(t)$ are found from $tx_n'(m) = 1$ and $tx_n'(m) + \beta tx_u'(m) = 1$, respectively. Also from the assumption of strict concavity, it is easy to see that $m_1'(t) > 0$, $m_2'(t) > 0$, and $m_2(t) > m_1(t)$ for all t. Since t represents the marginal rate of substitution of a type-t agent, I conclude that agents with higher preferences for car service select higher maintenance levels. Moreover, agents would also select higher maintenance levels for new cars if they anticipate using them in the second period.

Finally, assuming that the set of used car sellers who adopt the maintenance policy $m_1(t)$ is not empty, the average quality of traded used cars, $Ex_u(T_1)$ is given by

$$(6) \quad Ex_u(T_1) = \int_{T_1} x_u(m_1(t))w(t)\,dt / W(T_1)$$

where $W(T_1) = \int_{T_1} w(t)\,dt$ is the total supply of used cars by agents in T_1. The average quality of nontraded used cars can be similarly calculated for agents who keep new cars two periods (option 2). The only difference is that $m_2(t)$ is used for the calculation instead of $m_1(t)$.

II. Equilibrium in the Used Car Market

An equilibrium in the used car market can be defined as follows:

DEFINITION: *The used car market is in equilibrium, given P_n, if there exist x_u^e and P_u such that* (a) $x_u^e = Ex_u(T_1)$ *where* $T_1 \neq 0$, *and* (b) $W(T_1) = W(T_3)$.

[3] It is not difficult to incorporate such a possibility in the model by slightly modifying the definition of an equilibrium presented below. However, this would add no more significant result but complications. Moreover, such a case does not arise in two special cases on which I will concentrate later.

Condition (a) says that the expectations about the average quality of traded used cars must be correct. Condition (b) is the usual market-clearing condition in the Walrasian paradigm. Since $T_1 \neq 0$, there must be a positive level of trade in the market. A quick glance at the conditions reveals that the nature of equilibrium depends heavily on the shapes of the S_i curves, especially on their slopes, which are given below.

$$(7)\quad \begin{aligned} \partial S_1/\partial t &= \partial Z_1/\partial t \\ &= (1+\beta)x_n(m_1(t)) > 0 \\ \partial S_2/\partial t &= \partial Z_2/\partial t \\ &= x_n(m_2(t)) + \beta x_u(m_2(t)) > 0 \\ \partial S_3/\partial t &= \partial Z_3/\partial t = (1+\beta)x_u^e > 0 \\ \partial S_4/\partial t &= 0 \\ \partial^2 S_1/\partial t^2 &= (1+\beta)x_n'(m_1(t))m_1'(t) > 0 \\ \partial^2 S_2/\partial t^2 &= \{x_n'(m_2(t)) \\ &\quad + \beta x_u'(m_2(t))\}m_2'(t) > 0 \\ \partial^2 S_3/\partial t^2 &= \partial^2 S_4/\partial t^2 = 0. \end{aligned}$$

First note that the S_i's $(i=1,2,3,)$ are all increasing functions of t. Second, S_1 and S_2 are increasing at increasing rates while S_3 increases linearly. Third, S_1 is always steeper than S_3 because of the assumption that $x_n(0) > x_u(\infty)$. Fourth, it is not obvious whether or not S_2 is steeper than S_1 and S_3.

The last observation opens up an interesting possibility that there are different types of equilibrium depending on the set of agent types and the shape of $x_n(m)$ and $x_u(m)$. In this paper, I consider two special cases of interest: Case 1: S_2 is steeper than S_3 but flatter than S_1; and Case 2: S_2 is the steepest. These complete orderings make the analysis simple because the S_i curves intersect each other at most once. Even though I will be mainly concerned with the first case, the second case is equally interesting because it turns out to be very similar to the Akerlof-Wilson model in nature.

A

Case 1: $\partial S_1/\partial t > \partial S_2/\partial t > \partial S_3/\partial t$.

Case 1 is worth considering for several reasons. First, if $x_n(\infty)$ is bounded from above and t is sufficiently large, the above assumption is automatically satisfied, because as t grows, $x_n(m_2(t)) - x_n(m_1(t)$ becomes arbitrarily small while $x_n(m_1(t)) - x_u(m_2(t))$ is bounded away from some positive number so that S_1 will eventually be steeper than S_2 (see equation (7)). The same kind of reasoning can be applied to show that S_2 is steeper than S_3 for a sufficiently large t. Second, if the technology of $x_n(m)$ and $x_u(m)$ is of an exponential form, for which some numerical examples are constructed later on, it can be shown that S_1 is always steeper than S_2, which is in turn steeper than S_3. Finally and most importantly, this formulation gives us an interesting counterexample to Akerlof's Lemons Principle.

A possible equilibrium situation is drawn in Figure 1, assuming that all the T_i's are nonempty. Let t_1 be the marginal agents who are indifferent between options 1 and 2. Similarly, t_2 and t_4 are the marginal agents who are indifferent between options 2 and 3, and options 3 and 4, respectively. Then,

$$(5')\qquad T_1 = [t_1, \bar{t}], \quad T_2 = [t_2, t_1), \quad T_3 = [t_4, t_2); \quad T_4 = [\underline{t}, t_4).$$

From Figure 1, the following immediate observations can be made. Agents with higher preferences for car service (in T_1 and T_2) buy new cars and agents with lower preferences (in T_3) buy used cars. Among the agents buying new cars, agents with higher preferences (in T_1) are used car sellers and those with lower preferences (in T_2) are used car keepers.

It can be seen that, in equilibrium, used cars of various qualities are traded in the market. Agents with used cars of above average quality are also willing to sell their cars. In other words, as noted earlier, when an agent chooses the optimum maintenance level, he is only concerned with the net service from a car and not with the used car price,

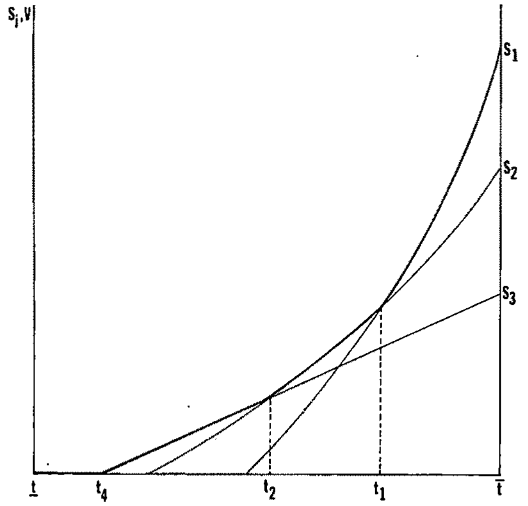


FIGURE 1

which is beyond his control under asymmetric information. That is, the fact that he receives a lower price than he possibly deserves is already taken into account in his decision process. Another direct result follows from the diagram.

PROPOSITION 1: *The average quality of nontraded used cars can be either higher or lower than that of traded used cars.*

This is because even though, for given t, $m_2(t) > m_1(t)$ so that $x_u(m_2(t)) > x_u(m_1(t))$, T_1 consists of agents with higher t's selecting higher maintenance levels for a given option as discussed in Section I. As a result, either one of them could be higher than the other. In some cases, the average quality of traded cars will be higher than that of nontraded cars, thus contradicting the Lemons Principle. A numerical example is provided below.

Numerical Example: Suppose that $x_n(m) = 2 - \exp(-m/2)$ and $x_u(m) = (1 - \exp(-m/2))/2$. Also t is distributed on $[7,8]$ such that $w(t) = .5$ for $7.65 < t < 7.90$, and equals 1 otherwise. If $P_n = 13.00$, then equilibrium t_i's are given by $t_1 = 7.76$, $t_2 = 7.56$, $t_4 = 7.39$, and $P_u = 2.76$. The average quality of nontraded cars is .44 while that of traded cars is .37, confirming the Lemons Principle. However, as expected, the result is very sensitive to the distribution of agent types (and also the technology). In particular, suppose the lower margin of a set of t over which $w(t) = .5$ increases slightly. Then the average quality of nontrade cars will be reduced with all other equilibrium values remaining unchanged. For example, if the lower margin goes up to 7.71 from 7.65, the average quality of nontrade cars drops to .32, becoming less than that of traded cars.

Now let us investigate the nature of equilibrium shown in Figure 1 more closely. In terms of t_i's ($i = 1,2,4$), the equilibrium can be characterized as

$$\text{(8a)} \quad Z_1(t_1) - Z_2(t_1) + \beta(-P_n + (1+\beta)P_u) = 0$$

$$\text{(8b)} \quad Z_2(t_2) - Z_3(t_2) + (-P_n + (1+\beta)P_u) = 0$$

$$\text{(8c)} \quad Z_3(t_4) - (1+\beta)P_u = 0$$

$$\text{(8d)} \quad \int_{t_1}^{\bar{t}} w(t)\,dt = \int_{t_4}^{t_2} w(t)\,dt$$

$$\text{(8e)} \quad \bar{t} > t_1 > t_2 > t_4 > \underline{t}.$$

I first derive the demand and supply functions in the used car market from the above equation. For this, I totally differentiate (8a), (8b), and (8c) with respect to P_u. Using the slope conditions and the fact that an increase in t_1 raises the average quality of traded used cars, it can easily be shown that

$$\text{(9)} \quad \partial t_1/\partial P_u < 0, \quad \partial t_2/\partial P_u < 0, \quad \partial t_4/\partial P_u > 0.$$

Combined with the market-clearing condition equation (8d), the above equation says that the demand function is downward sloping while the supply function is upward sloping. Therefore,

PROPOSITION 2: *If there exists an equilibrium, it is unique.*

The reason why the demand and supply functions have the usual shape is the following. Suppose that the used car price goes up. This will make it more profitable than before to sell used cars, which will induce some agents who otherwise would be used car keepers to become used car sellers, lowering the average quality and increasing the supply of used cars. At the same time, an increase in the used car price and the resulting decrease in the average quality will depress the demand for used cars.

Before continuing, it seems worthwhile to investigate the structural difference between the present model and the Akerlof-Wilson model. In the Akerlof-Wilson model, sellers are identical except for the quality of cars they own. Then, given a used car price, owners of cars above some critical quality level do not want to offer their cars for sale in the market. So identical agents behave in different ways depending on the cars they own. By contrast, in this model, there are many different types of agents. All agents of the same type are assumed to behave in the same way, that is, take the same option so that the quality of a car is endogenously linked to the type of agent holding the car. This structural difference generates different outcomes in many ways. For example, in case of an increase in the used car price, the Akerlof-Wilson model predicts an opposite result that the quality will go up because owners of higher quality cars will be induced to sell.

There are a few more things to be noted. First, if $\underline{t} = \bar{t}$ (the agents are identical so that buyers get perfect information about the quality of used cars), a used car market cannot exist. This is because, in such a situation, for a used car market to exist, the typical agent must be indifferent between options 1 and 3. But this means that the typical agent in turn is indifferent between those options and the option of buying a new car, maintaining at the level of m_1, and keeping it for both periods. However, the latter, and therefore options 1 and 3, are obviously inferior to option 2, that constitutes a desired contradiction. As a result, in this case, the agents will keep their used cars because they cannot obtain gains from trade. This observation can be easily extended to a case where $\underline{t}$ and $\bar{t}$ are close to each other. That is, if the difference in consumer tastes is small enough, there may not exist an equilibrium in the used car market.

Second, the effect of an increase in the new car price is examined. For this, I totally differentiate equations (8a), (8b), (8c), and (8d) with respect to P_n. After tedious manipulation, we have

$$(10) \quad \partial t_1/\partial P_n > 0, \quad \partial t_2/\partial P_n > 0,$$
$$\partial t_4/\partial P_n > 0, \quad 1/(1+\beta) > \partial P_u/\partial P_n > 0.$$

A type-t_1 agent who was indifferent between options 1 and 2 will prefer option 2. This is because as the new car price goes up, the additional cost is $(1+\beta)(dP_n - \beta\, dP_u)$ which is greater than dP_n, the additional cost if he takes option 2, by the second last inequality of equation (10). An analogous argument can be made for agents of types t_2 and t_4. Therefore, as the price of new cars increases, the supply of and demand for used cars decrease and, as a result, the average quality of used cars rises. The used car price also rises, but by a relatively small amount. To see the effect on agents's welfare, note that the S_1 and S_2 curves shift downward. The S_3 curve becomes steeper because of an increase in the average quality. This observation combined with equation (10) indicates that $\partial S_3/\partial P_n$ may be positive for some t's around t_2, reflecting that the extra utility from the average quality increase exceeds the disutility from the used car price increase for those agents. Therefore, everyone is made worse off in the new equilibrium except possibly some used car buyers (around type-t_2 agents) who benefit from the increased average quality.

B

Case 2: $\partial S_2/\partial t > \partial S_1/\partial t$.

Case 2 is shown as a solid line in Figure 2 where $T_1 = [t_3, t_1)$, $T_2 = [t_1, \bar{t}]$, $T_3 = [t_4, t_3)$, and $T_4 = [\underline{t}, t_4)$. It is immediately seen that in this case the average quality of nontraded cars is definitely higher than that of traded

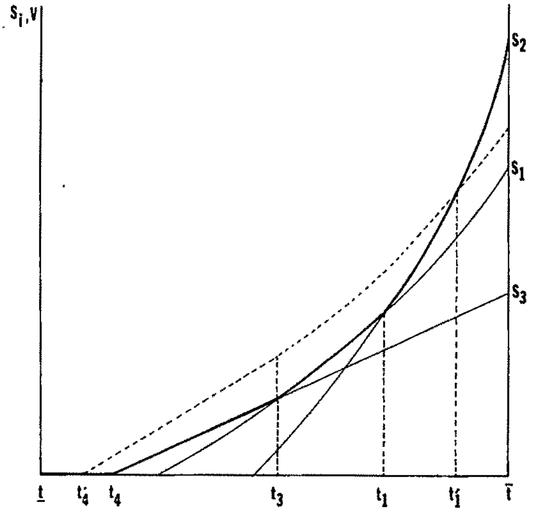


FIGURE 2

cars, thus supporting the Lemons Principle. Therefore, in this special case, the present model and the Akerlof-Wilson model give the same conclusion. However, there still exists a significant discrepancy between the two models, illustrated by the following example. Wilson argues that in the Walrasian paradigm, there may exist multiple equilibria, and in such a case an equilibrium with the highest used car price is Pareto superior. Multiple equilibria could also exist in Case 2 for the same reason given by Wilson. To see this point more closely, suppose that P_u increases. This will induce the agents with higher t's than t_1 to sell their cars, which will increase the average quality of traded cars. At the same time, this increase in P_u may attract some used car buyers to be used car sellers, which has a negative effect on the average quality. As Wilson argues, if the average quality happens to increase more rapidly than the used car price, there might be another equilibrium with a higher used car price.

If t_3 does not change in the new equilibrium (which may be interpreted as saying that sellers and buyers are separated as in Akerlof-Wilson model), then Wilson's argument can be applied in exactly the same way to conclude that the new equilibrium is Pareto superior. This possibility is shown in Figure 2 where a dotted line is an envelope of new S_1 and S_3 curves. The supply of used cars increases because $t_1' > t_1$. Then the demand should also increase, which implies that S_3 cuts S_4 before t_4 in the new equilibrium. Obviously, agents in $[t_4', t_1')$ are better off so that the new equilibrium is Pareto superior.

But, in general, it will be more reasonable to think that t_3 changes after the price change. Then, there might be an equilibrium with t_3 and t_4 moving to the right which clearly cannot be Pareto superior because some used car buyers will be driven out of the market due to the increase in P_u.

PROPOSITION 3: *In Case 2, the Lemons Principle holds. Moreover, there are possibly multiple equilibria. However, it is not generally possible to rank those equilibria by the Pareto criterion.*

III. Concluding Remarks

I have constructed a model of the used car market with asymmetric information where there are many types of agents in the market. Even though there could be many kinds of equilibrium depending on the underlying parameters, attention has been focused on two contrasting cases.

In Case 1, there exists a unique equilibrium given the new car price and the distribution of agent types. Furthermore, the Lemons Principle need not hold: average quality of traded used cars may be higher than that of nontraded cars. Case 2 closely resembles the Akerlof-Wilson model. Not only does the Lemons Principle hold in this case, but also multiple equilibria with different used car prices may emerge for the reason Wilson indicates. However, in this case, the ability of agents to change their positions costlessly makes it impossible to rank equilibria by the Pareto criterion in contrast with Wilson's argument.

In future research, it may be worthwhile to consider the situation where perfect information is available; that is, each agent may learn the quality of used cars by observing them. The importance of such a study may

be seen by the following consideration. In reality, a potential buyer of a used car buys it either from an individual seller or from a used car dealer. In the first type of transaction, the buyer presumably obtains little information about the quality of the car, which fits the present model. On the other hand, if the buyer is engaged in the second type of transaction, he may obtain considerable information from the used car dealer, for example, in the form of a warranty, by paying a transaction cost. Perfect information may be accommodated by modifying the model so that an agent has perfect knowledge about the quality of a car if he pays some positive cost. It will be necessary to combine the two models so as to allow each agent to decide on what types of transactions he will make in order to explain why both types of transactions are popular in the real world.

REFERENCES

Akerlof, George A., "The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, August 1970, *84*, 488–500.

Bond, Eric W., "A Direct Test of the "Lemons" Model: The Market for Used Pickup Trucks," *American Economic Review*, September 1982, *72*, 836–40.

Wilson, Charles, "The Nature of Equilibrium in the Markets with Adverse Selection," *Bell Journal of Economics*, Spring 1980, *11*, 108–30.

A Note on the General Validity of the Heckscher-Ohlin Theorem

By HELMUT FORSTNER*

In his article in this *Review* (1982) Alan Deardorff proved, among other results, a generalization of the commodity version of the Heckscher-Ohlin (H-O) Theorem. In the formulation of this general hypothesis, Deardorff used the unconventional concept of a generalized covariance, named *comvariance* (*com*), among three variables. In particular, he proved that $com(x, y, z)$ is positive, where x, y, and z measure a country's factor abundance, a good's factor intensity and a country's net exports, respectively. In simple language, this result shows that *on average* countries will be net exporters of those goods whose production is intensive in the countries' abundant factor(s).[1] While the rather artificial construct of *com* treats the measures of factor endowment, factor intensity, and trade as fully symmetrical, most empirical tests of the H-O hypothesis are based on regression models that relate a dependent trade variable to independent factor intensity variables.

This note demonstrates that Deardorff's result can be extended to yield *weak* hypotheses within the framework of linear regression analysis. I prove two corollaries to Deardorff's theorem on the factor content of trade that yield statements about the signs, on average, of regression coefficients. This extension may help clarify some of the confusion that has arisen regarding the question of the theoretical validity of regression tests of the factor proportions theory.[2]

I. Assumptions

The assumptions underlying Deardorff's model can be summarized as follows. There are l internationally nontradeable factors ($l \geq 2$) and m tradeable goods ($m \geq l$) in n countries which have identical technologies (represented by a set $\mathscr{K}$ which has all the properties necessary for the existence of the equilibria studied). For each country, domestic demand is homogeneous of degree zero in prices and income, and satisfies the weak axiom of revealed preference. Further assumptions are those of perfect competition, constant returns to scale and the absence of external economies, production taxes, and similar distortions. A country-autarky equilibrium consists of a combination of net outputs, prices (of goods), and wages (of factors) that are feasible, demanded, and profit maximizing. A world-trade equilibrium is defined as a combination of world prices, domestic prices, and quantities produced, traded, and demanded by each country, so that production and trade are feasible, demands are met, and profits are maximized. For each country it is assumed in addition that its trade balance at world prices is zero, and that eventual differences between world and domestic prices are the result, on average, of taxes on trade rather than of subsidies. For each good the world market is cleared.

In the discussion below, the following definitions and designations will be used.

*Division for Industrial Studies, United Nations Industrial Development Organization (UNIDO), P.O. Box 400, Vienna International Centre, A–1400 Vienna, Austria. The views expressed in this paper do not necessarily reflect those of the United Nations. I am indebted to Tracy Murray for detailed comments on an earlier draft and to an anonymous referee for valuable suggestions concerning in particular the extension of the original results. I have also benefited from discussions with Robert H. Ballance.

[1] This on average result will be referred to as a *weak* formulation of the H-O hypothesis. The strong commodity version of the above hypothesis has been proved only for the two-factor, two-good, two-country world.

[2] See, for example, the critical comment on Jon Harkness (1978) by Edward Leamer and Harry Bowen (1981).

The value of net exports of good i by country j at the world price p_i^w is

$$\tau_{ij} = p_i^w T_{ij}; \tag{1}$$

and

$$\omega_{hj} = (\bar{w}_h - w_{hj}^a)/\bar{w}_h \tag{2}$$

is a measure of the abundance of factor h in country j, where w_{hj}^a is the autarky price of factor h in country j and

$$\bar{w}_h = \frac{1}{n}\sum_{j=1}^{n} w_{hj}^a; \tag{3}$$

$$\vartheta_{hij} = \bar{w}_h g_{hij}^t / p_i^w \tag{4}$$

is a measure of factor intensity for factor h based on actual factor requirements of trade $g_{hij}^t (1 \le h \le l, 1 \le i \le m, 1 \le j \le n)$ (as described in Deardorff, p. 687);

$$z_{hij} = \vartheta_{hij}\omega_{hj} + A \tag{5}$$

is a nonnegative measure of the interaction[3] between country j's endowment with factor h and the h-intensity of j's trade in good i, where A is a constant with

$$A \ge |\min(\vartheta_{hij}\omega_{hj})| \tag{6}$$

with $z_{hij} \ge 0$ for all h, i, and j.

The following $(m \times 1)$ vectors and $(m \times l)$ matrices corresponding to the variables defined above will be used: $\tau_j =$ country j's net exports; z_{hj}: country j's (intensity-endowment) interactions with respect to factor h; $\mathbf{i} = (1,1,\dots,1)'$; and $\mathbf{Z}_j = (\mathbf{z}_{1j}, \mathbf{z}_{2j}, \dots, \mathbf{z}_{lj})$ = country j's interactions matrix.

In addition $\|x\|$ will denote the Euclidean norm in m-dimensional real space and

$$\bar{\mathbf{x}} = \frac{1}{l}\sum_{h=1}^{l} \mathbf{x}_h$$

will designate the centroid of l points $\{\mathbf{x}_h\}$ in this space. The cosine between two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined by $\cos(\mathbf{x},\mathbf{y}) = \mathbf{x}'\mathbf{y}/\|\mathbf{x}\|\,\|\mathbf{y}\|$.

II. Two Corollaries to Deardorff's Theorem on the Factor Content of Trade

Country-specific interactional regression tests are based on relationships of the general form

$$\tau_j = f(\mathbf{Z}_j, \boldsymbol{\beta}_j) + \mathbf{u}_j, \tag{7}$$

where $\boldsymbol{\beta}_j$ is a vector of unknown parameters and $\mathbf{u}_j$ a vector of random disturbances. As an extension of Corollary 2 in Deardorff (p. 691), a general version of the H-O Theorem can be stated in the form of a hypothesis about the sign of a weighted average of the least squares parameter estimates of the country models (7). The two corollaries presented in the following deal with two different specifications of the regression model (7). Corollary A applies to the case of regressing net exports on individual columns of the interactions matrix, whereas Corollary B states a result about regression (7) in the absence of linear restrictions on parameters.

COROLLARY A: *For each factor h $(1 \le h \le l)$ and each country j $(1 \le j \le n)$, let an interactional regression equation be specified by*

$$\tau_j = \mathbf{i}\alpha_{hj} + \mathbf{z}_{hj}\beta_{hj} + \mathbf{u}_{hj}, \tag{8}$$

where $\mathbf{u}_{hj}$ is the disturbance term and τ_j and $\mathbf{z}_{hj}$ are as defined above. If the ordinary least squares (OLS) procedure is applied to each equation separately, the cross-commodity estimates $\hat{\beta}_{hj}$ satisfy (under the assumptions stated previously) the following average relationship:

$$\sum_{h=1}^{l}\sum_{j=1}^{n} s_{hj}\hat{\beta}_{hj} > 0, \tag{9}$$

with

$$s_{hj} = \frac{1}{m}\sum_{i=1}^{m}(z_{hij} - \bar{z}_{hj})^2. \tag{10}$$

[3]For a discussion of the concept of an interactional regression model of trade, see, for example, Roger Bowden (1983, pp. 216–18). The translation of the interaction measure to the nonnegative part of the real line will yield a proper weight variable in Corollary B below.

PROOF:

The expression for the *OLS* parameter estimate

$$(11)\quad \hat{\beta}_{hj}=\frac{m\sum_{i=1}^{m} z_{hij}\tau_{ij}-\sum_{i=1}^{m} z_{hij}\sum_{i=1}^{m}\tau_{ij}}{m\sum_{i=1}^{m}(z_{hij})^2-\left(\sum_{i=1}^{m} z_{hij}\right)^2}$$

can be rewritten as

$$(12)\qquad \hat{\beta}_{hj}=\sum_{i=1}^{m} z_{hij}\tau_{ij}/m\cdot s_{hj},$$

due to the assumption of balanced trade for each country. This leads to

$$(13)\quad m\sum_{h=1}^{l}\sum_{j=1}^{n} s_{hj}\hat{\beta}_{hj}=\sum_{h=1}^{l}\sum_{i=1}^{m}\sum_{j=1}^{n} z_{hij}\tau_{ij}$$

$$=\sum_{h=1}^{l}\sum_{i=1}^{m}\sum_{j=1}^{n}(\vartheta_{hij}\omega_{hj}+A)\tau_{ij}$$

$$=\sum_{h=1}^{l}\sum_{i=1}^{m}\sum_{j=1}^{n}\omega_{hj}\vartheta_{hij}\tau_{ij}>0,$$

where the last equality follows again from balanced trade and the inequality from relationship (46) in Deardorff (p. 691).

COROLLARY B: *For each country j $(1\le j\le n)$, let an interactional regression model be specified by*

$$(14)\qquad \boldsymbol{\tau}_j=\mathbf{Z}_j\boldsymbol{\gamma}_j+\mathbf{u}_j,$$

where $\boldsymbol{\gamma}_j$ is the parameter vector and $\mathbf{u}_j$ is the disturbance term.[4] *Under the assumptions outlined above the following relationship holds for the least squares estimates $\hat{\boldsymbol{\gamma}}_j$ obtained from cross-commodity estimation for each j separately*:

$$(15)\qquad \sum_{h=1}^{l}\sum_{j=1}^{n} t_{hj}\hat{\gamma}_{hj}>0$$

with

$$(16)\quad t_{hj}=\|\mathbf{z}_{hj}\|\,\|\bar{\mathbf{z}}_j\|\cos(\mathbf{z}_{hj},\bar{\mathbf{z}}_j)\ge 0.$$

PROOF:

For country j's regression the normal equations can be written

$$(17)\qquad \mathbf{Z}_j'\mathbf{Z}_j\hat{\boldsymbol{\gamma}}_j=\mathbf{Z}_j'\boldsymbol{\tau}_j,$$

or

$$(18)\quad \sum_{k=1}^{l}\left(\mathbf{z}_{hj}'\mathbf{z}_{kj}\right)\hat{\gamma}_{kj}=\mathbf{z}_{hj}'\boldsymbol{\tau}_j\qquad(1\le h\le l).$$

Summation over h yields

$$(19)\quad \sum_{h=1}^{l}\sum_{k=1}^{l}\left(\mathbf{z}_{hj}'\mathbf{z}_{kj}\right)\hat{\gamma}_{kj}=\sum_{h=1}^{l}\mathbf{z}_{hj}'\boldsymbol{\tau}_j,$$

or, equivalently,

$$(20)\qquad l\sum_{k=1}^{l}\left(\bar{\mathbf{z}}_j'\mathbf{z}_{kj}\right)\hat{\gamma}_{kj}=\sum_{h=1}^{l}\mathbf{z}_{hj}'\boldsymbol{\tau}_j.$$

Summation over j in turn leads to

$$(21)\quad l\sum_{h=1}^{l}\sum_{j=1}^{n}\left(\bar{\mathbf{z}}_j'\mathbf{z}_{hj}\right)\hat{\gamma}_{hj}=\sum_{h=1}^{l}\sum_{j=1}^{n}\mathbf{z}_{hj}'\boldsymbol{\tau}_j$$

$$=\sum_{h=1}^{l}\sum_{i=1}^{m}\sum_{j=1}^{n} z_{hij}\tau_{ij}>0,$$

where the inequality follows from the proof of Corollary A. Thus, finally the inequality

$$(22)\qquad \sum_{h=1}^{l}\sum_{j=1}^{n}\left(\bar{\mathbf{z}}_j'\mathbf{z}_{hj}\right)\hat{\gamma}_{hj}>0$$

[4] Whenever factor shares are used to define factor intensities and production functions are assumed linearly homogeneous, a zero intercept is suggested for equations like (14) for the simple reason of guarding against perfect multicollinearity.

must hold which entails (15) by virtue of

$$\bar{\mathbf{z}}_j'\mathbf{z}_{hj} = \|\bar{\mathbf{z}}_j\|\,\|\mathbf{z}_{hj}\|\cos(\bar{\mathbf{z}}_j,\mathbf{z}_{hj}). \tag{23}$$

The essence of the above two corollaries is that *OLS* parameter estimates from interactional regression tests of the H-O hypothesis must (under the given assumptions) be *positive* on average. While Corollary A refers to "total" regression coefficients of the interactions between factor intensities and factor endowments, Corollary B deals with the corresponding "partial" coefficients. In both cases it is demonstrated that a weighted average of the interaction coefficients across all factors and all countries is positive. For each total regression coefficient $\hat{\beta}_{hj}$, the weight is given by the sample variance s_{hj} across all m commodities of the interactions z_{hij}. Each partial regression coefficient $\hat{\gamma}_{hj}$ has to be weighted by an expression which depends on the "strength" of h-specific interactions (via $\|\mathbf{z}_{hj}\|$), the overall strength of interactions in country j (via $\|\bar{\mathbf{z}}_j\|$) and the position of $\mathbf{z}_{hj}$ with respect to the interaction centroid $\bar{\mathbf{z}}_j$ (via cos $(\mathbf{z}_{hj},\bar{\mathbf{z}}_j)$).

III. Conclusions

Regression tests of the H-O Theorem usually suffer from the limitation of the traditional commodity version to the abstract case of two factors, two goods, and two countries.[5] Deardorff's general form of the theorem represents a weaker version of the factor proportions hypothesis stating that, *on average*, countries will be net exporters of those goods that are intensive in the use of abundant factors of production. In the present paper, a corresponding weak regression hypothesis has been derived, which simply states that, on average, the *OLS*-estimated coefficients of interactional regression models will carry the (intuitively expected) *positive* signs. This weak hypothesis may narrow the gap between the algebraic proofs of formal theory and the numerous attempts to test the H-O proposition by use of regression procedures.

It has to be borne in mind, however, that the present results are limited in their empirical applicability. These limitations have to do with the measurability of the variables included in the above regressions as well as in Deardorff's theoretical proposition. First, the measure of factor abundance is based on autarky factor prices which cannot be observed. Second, the factor-intensity variable involves autarky factor prices, actual factor requirements of trade, and world goods prices, and must thus be expected to differ in general from observable factor shares. Moreover, a massive amount of detailed information about production techniques in different countries would be required to construct matrices of actual total factor requirements of each country's trade. Third, the averages of regression coefficients in the two corollaries have to be taken across all countries and all factors, whereas empirical studies are usually limited (due to the lack of data) to subsets of countries and factors.

The limitations listed above become less severe, if regression hypotheses are derived from theoretical propositions that correspond more closely to the empirical reality than Deardorff's general version of the H-O Theorem. A case in point is a result on the factor content of trade stated recently by Elhanan Helpman (1984). Let free-trade equilibrium factor prices replace autarky factor prices in the definitions of factor abundance and of factor intensity. Then, a slightly weaker version of Corollary B can be obtained, if factor prices are not equalized among countries and a number of fairly standard assumptions (including those of no intermediate goods and of balanced trade for each country) are fulfilled. In the Appendix, Helpman's results are used to derive the inequality

$$\sum_{h=1}^{l}\sum_{j=1}^{n} t_{hj}\hat{\gamma}_{hj} \geq 0, \tag{24}$$

where the notation of Corollary B is em-

[5] Bee-Yan Aw (1983) demonstrated that the regression procedure used in most empirical studies may be inappropriate as a test of the factor proportions theory, if more than two factors of production are involved.

ployed. The above-stated relationship is seen to be less limited in an empirical sense than its counterpart (15), because it involves post-trade data on factor abundance. However, the factor-intensity measure used still includes actual factor requirements of trade.

Furthermore, Helpman's restriction on the factor content of bilateral trade—that likewise holds in the absence of factor price equalization—can be reformulated as an average relationship between regression coefficients. Let τ_{jk} be the vector of bilateral net exports of country j to country k valued at world prices, $\mathbf{Z}_j^{(jk)}$ the matrix of interactions between factor intensities of τ_{jk} (based on actual factor requirements) and factor endowments of j (measured by use of post-trade factor rewards as indicated above), and $\mathbf{Z}_k^{(jk)}$ the analogous interactions matrix for country k. Then, for the *OLS* parameter estimates of the two regressions

$$(25) \qquad \tau_{jk} = \mathbf{Z}_j^{(jk)}\gamma_j + \mathbf{v}_j$$

$$(26) \qquad \tau_{jk} = \mathbf{Z}_k^{(jk)}\gamma_k + \mathbf{v}_k$$

the inequality

$$(27) \qquad \sum_{h=1}^{l} t_{hj}\hat{\gamma}_{hj} \geq \sum_{h=1}^{l} t_{hk}\hat{\gamma}_{hk}$$

holds under the assumptions made in Section III of Helpman's article. The weights t_{hj} and t_{hk} $(h=1,2,\ldots,l)$ in (27) are constructed from the interactions matrices $\mathbf{Z}_j^{(jk)}$ and $\mathbf{Z}_k^{(jk)}$ in the way described in connection with Corollary B.[6] A simple interpretation of inequality (27) would be that bilateral net exports are on average not "explained better" by the factor-intensity/factor-endowment interactions of the importer than by those of the exporter. More loosely interpreted, bilateral net exports reflect the interactions of the exporter rather than those of the importer. This proposition about bilateral trade shares with the modified version of Corollary

[6]For a proof of the above relationship, see the Appendix.

B the advantages with respect
work.

Finally, a general reservatioı made which applies to the re potheses stated in this note as w negative results on regression H-O theory presented by Leamer and by Aw. It concerns the me limitations that arise from the f rivations are confined to the parameter estimation. Although– to expected future extensions—t ogy of stochastic regression mod used in formulating the corollar dorff's findings, the results of note should be seen as pertai realm of descriptive regression a deeper-lying specification proble in modeling the relationship be flows, factor intensities, factor e and other elements that might trade have not been treated in Nevertheless, it would be of inte the above-stated corollaries can to hold for the interactional parameters of more complete mo

APPENDIX

This Appendix proves the val inequalities (24) and (27).

For a proof of relationship (2 ficient to demonstrate that the $\sum_{h=1}^{l}\sum_{i=1}^{m}\sum_{j=1}^{n} z_{hij}\tau_{ij}$ is nonnega z_{hij} is based on free-trade equilil prices w_{hj}^{f}. From the definitions i the assumption of balanced cour follows that

$$(A1) \quad \sum_{h=1}^{l}\sum_{i=1}^{m}\sum_{j=1}^{n} z_{hij}\tau_{ij}$$

$$= \sum_{h=1}^{l}\sum_{i=1}^{m}\sum_{j=1}^{n} (\omega_{hj}$$

$$= \sum_{h=1}^{l} \bar{w}_h^f \sum_{i=1}^{m}\sum_{j=1}^{n} g_h$$

$$- \sum_{h=1}^{l}\sum_{i=1}^{m}$$

where g_{hij} is the actual requirement of factor h per unit of net exports T_{ij} and $\bar{w}_h^t = \sum_{j=1}^n w_{hj}^t/n$. The first of the last two terms in (A1) is zero, since it is the overall sum of actual factor contents of trade (see Deardorff, p. 687). To evaluate the last term, relationship (12) of Helpman's paper can be rewritten to obtain

$$\text{(A2)} \quad \sum_{i=1}^m p_i^w T_{ij} \geq \sum_{h=1}^l w_{hj}^t \sum_{i=1}^m g_{hij} T_{ij}.$$

Then the desired inequality

$$\text{(A3)} \quad \sum_{h=1}^l \sum_{i=1}^m \sum_{j=1}^n w_{hj}^t g_{hij} T_{ij} \leq 0$$

follows from the assumption of balanced trade.

The validity of relationship (27) can also be demonstrated on the basis of Helpman's results and the technique of Corollary B. Application of the latter to the regressions (25) and (26) leads to

$$\text{(A4)} \quad l \sum_{h=1}^l t_{hj}\hat{\gamma}_{hj} = \sum_{h=1}^l \bar{w}_h^t \sum_{i=1}^m g_{hi(jk)} T_{i(jk)} - \sum_{h=1}^l w_{hj}^t \sum_{i=1}^m g_{hi(jk)} T_{i(jk)} + Al \sum_{i=1}^m p_i^w T_{i(jk)}$$

and

$$\text{(A5)} \quad l \sum_{h=1}^l t_{hk}\hat{\gamma}_{hk} = \sum_{h=1}^l \bar{w}_h^t \sum_{i=1}^m g_{hi(jk)} T_{i(jk)} - \sum_{h=1}^l w_{hk}^t \sum_{i=1}^m g_{hi(jk)} T_{i(jk)} + Al \sum_{i=1}^m p_i^w T_{i(jk)}$$

where $g_{hi(jk)}$ represents the actual requirement of factor h per unit of good i's net exports $T_{i(jk)}$ from country j to country k.

Hence it follows that

$$\text{(A6)} \quad \sum_{h=1}^l t_{hj}\hat{\gamma}_{hj} - \sum_{h=1}^l t_{hk}\hat{\gamma}_{hk} = \frac{1}{l} \sum_{h=1}^l \left(w_{hk}^t - w_{hj}^t\right) \sum_{i=1}^m g_{hi(jk)} T_{i(jk)},$$

where the last expression is seen to be nonnegative on the basis of Helpman's inequality (18) which states that the value at country j's factor prices of net exports from country j to country k of trade embodied factors is not greater than the corresponding value at country k's factor prices.

REFERENCES

Aw, Bee-Yan, "The Interpretation of Cross-Section Regression Tests of the Heckscher-Ohlin Theorem with Many Goods and Factors," *Journal of International Economics*, February 1983, *14*, 163–67.

Bowden, Roger J., "The Conceptual Basis of Empirical Studies of Trade in Manufactured Commodities: A Constructive Critique," *Manchester School of Economic and Social Studies*, September 1983, *51*, 209–34.

Deardorff, Alan V., "The General Validity of the Heckscher-Ohlin Theorem," *American Economic Review*, September 1982, *72*, 683–94.

Harkness, Jon, "Factor Abundance and Comparative Advantage," *American Economic Review*, December 1978, *68*, 784–800.

Helpman, Elhanan, "The Factor Content of Foreign Trade," *Economic Journal*, March 1984, *94*, 84–94.

Leamer, Edward E. and Bowen, Harry P., "Cross-Section Tests of the Heckscher-Ohlin Theorem: Comment," *American Economic Review*, December 1981, *71*, 1040–43.

Screening vs. Rationing in Credit Markets with Imperfect Information

By HELMUT BESTER*

Will credit be rationed in markets with imperfect information? Credit rationing is said to occur when some borrowers receive a loan and others do not, although the latter would accept even higher interest payments or an increase in the collateral.[1] In their 1981 article, Joseph Stiglitz and Andrew Weiss argued that banks may prefer to reject some borrowers because of negative adverse selection and incentive effects: for a given collateral, an increase in the rate of interest causes adverse selection, since only borrowers with riskier investments will apply for a loan at a higher interest rate. Similarly, higher interest payments create an incentive for investors to choose projects with a higher probability of bankruptcy. On the other hand, for a fixed rate of interest, an increase in collateral requirements may result in a decline of a bank's profits as well. Stiglitz and Weiss show that this happens if the more risk-averse borrowers, who choose relatively safe investment projects, drop out of the market. Hildegard Wette (1983) has shown that increasing collateral requirements may result in adverse selection even with risk-neutral investors.

The purpose of this paper is to show that no credit rationing will occur in equilibrium if banks compete by choosing collateral requirements and the rate of interest to screen investors' riskiness. This argument relies on the assumption that banks decide upon the rate of interest and the collateral of their credit offers simultaneously rather than separately. Therefore, it becomes possible to use different contracts as a self-selection mechanism.[2] It is shown that investors with a low probability of bankruptcy are more inclined to accept an increase in collateral requirements for a certain reduction in the rate of interest than those with a high probability of failure.

Credit market equilibrium under imperfect information about investors' riskiness is analyzed in Section II. It is shown that, in equilibrium, no borrower is denied credit. The intuition is as follows. If some investor with a low probability of repayment does not receive the loan he prefers, then he will apply also for those contracts that are chosen by less risky borrowers. Therefore, a credit rationing equilibrium always pools good and bad risks. However, pooling of different risks at one contract is not viable against competition whenever self-selection mechanisms are available. If pooling occurs, then there exists another credit offer that is profitable because it attracts only the good risks from the pooling contract. An equilibrium is characterized by separation of borrowers of different risk. Borrowers with high probability of default choose a contract with a higher interest rate and a lower collateral than borrowers with low probability of default.

I. The Model

Consider a credit market with two types of entrepreneurs or firms, $i = a, b$. Each en-

*University of Bonn, Wirtschaftstheoretische Abtlg. III, Adenauerallee 24–26, 53 Bonn 1, West Germany. I thank Martin Hellwig for helpful suggestions and comments on an earlier draft. The final version was written while I was visiting the Department of Economics, University of California-San Diego. Financial support from the Deutsche Forschungsgemeinschaft is gratefully acknowledged.

[1]This is, of course, a different notion of rationing than, for example, Dwight Jaffee and Thomas Russell's: "Credit rationing occurs when lenders quote an interest rate on loans and then proceed to supply a smaller loan size than demanded by the borrowers" (1976, p. 651).

[2]A similar argument can be developed if investors can choose among a variety of projects with different riskiness. Collateral requirements can serve as an incentive mechanism because a higher collateral enforces a selection of less risky projects (see my 1984a paper).

trepreneur has the opportunity to undertake a project which requires a fixed amount of investment I.[3] The return to firm i's project is given by a random variable $0 \leq \tilde{R}_i \leq \bar{R}_i$ with distribution function $F_i(R)$. As in Stiglitz and Weiss, it will be assumed that $\tilde{R}_b$ corresponds to greater risk than $\tilde{R}_a$ in the sense of a mean-preserving spread,[4] such that

$$(1)\quad E\{\tilde{R}_a\} = E\{\tilde{R}_b\},$$

$$\int_o^y [F_b(R) - F_a(R)]\, dR \geq 0,$$

for all $y \geq 0$. Furthermore, let $F_i(R) > 0$ for all $R > 0$,[5] $i = a, b$. There are N_i entrepreneurs of type i.

Entrepreneurs have an initial wealth endowment of $W < I$. They finance their projects by borrowing the amount $B = I - W$. Given the loan size B, a credit contract $\gamma = (r, C)$ is specified by the rate of interest r and the collateral C charged by the bank. Entrepreneurs may face costs to collateralization. For simplicity, these costs will be assumed to be proportional to the amount of collateral by a factor $k \geq 0$. Firm i is said to be bankrupt if $C + R_i < (1+r)B$. If this happens, the bank becomes the owner of the investment project and its returns. Therefore, the expected profits of firm i by undertaking the project under a credit contract γ are given by

$$(2)\quad \Pi_i(\gamma) = E\{\max[\tilde{R}_i - (1+r)B - kC, -(1+k)C]\}.$$

On a loan γ to firm i the bank receives the expected rate of return

$$(3)\quad \rho_i(\gamma) = E\{\min[(1+r)B, \tilde{R}_i + C] - B\}/B.$$

Only contracts with $C \leq (1+r)B$ will be considered.[6] Banks finance their credit offers by funds from depositors. If π is the interest rate paid on deposits, then the bank's net profits on a loan γ to firm i are given by $[\rho_i(\gamma) - \pi]B$. The supply of loanable funds to the banking system will be described by a function $L^s(\cdot)$ of π. Let L^s be continuous, strictly increasing and let $L^s(0) = 0$.

Banks are unable to distinguish borrowers of different risk directly. They can do so only by offering a pair $(\gamma_\alpha, \gamma_\beta)$ of different credit contracts that act as a self-selection mechanism. The pair $(\gamma_\alpha, \gamma_\beta)$ is said to be *incentive compatible* if

$$(4)\quad \Pi_a(\gamma_\alpha) \geq \Pi_a(\gamma_\beta);\ \Pi_b(\gamma_\beta) \geq \Pi_b(\gamma_\alpha).$$

It is assumed that banks act as perfect competitors, that is, each bank takes the rate of interest π on deposits and the set of credit offers by competing banks as given and as independent of its own actions.

Entrepreneur i will invest only if he receives a loan γ such that $\Pi_i(\gamma) \geq (1+\pi)W$. Whenever a pair $(\gamma_\alpha, \gamma_\beta)$ of contracts is offered, he prefers the contract which maximizes his expected profits.[7] Hence, given the number of investors of each type and their decision rule, the demand for credit depends upon the set of available loan offers $\{\gamma_\alpha, \gamma_\beta\}$ and may be written as $L^d(\gamma_\alpha, \gamma_\beta)$.

II. Credit Market Equilibrium

Whether a particular entrepreneur is willing to undertake a project or not depends on the set of credit contracts offered by the banking system. Figure 1 shows the indifference curves[8] of each type i of firms for all contracts γ such that $\Pi_i(\gamma) = (1+\pi)W$. Below b, b', all b investors wish to obtain a loan while type a firms do so below a, a'. As

[3] Since the amount of investment is taken to be fixed, it cannot be used to convey information about a borrower's default risk. For models in which the level of investment is used as a signal for the repayment probability of borrowers, see my 1984b paper, and Hellmuth Milde and John Riley (1984).

[4] See Michael Rothschild and Stiglitz (1970).

[5] This condition ensures that there is a positive probability of default whenever the interest payments of a credit contract exceed the collateral.

[6] If $C > (1+r)B$, then a firm would never admit to being bankrupt. Hence, it is assumed that bankruptcy becomes observable only after a firm declares to be bankrupt.

[7] The case of risk-averse entrepreneurs is studied in my 1984a paper.

[8] Notice that investors' indifference curves will in general not be concave.

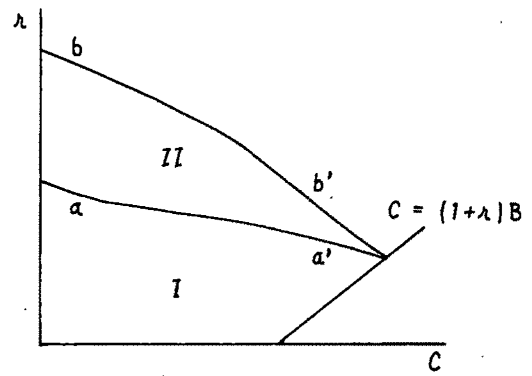


FIGURE 1

shown in Figure 1, one has

$$(5)\quad \Pi_b(\gamma) > \Pi_a(\gamma) \quad \text{if} \quad C < (1+r)B,$$

because firm i's profits are a convex function of $\tilde{R}_i$. As was observed by Stiglitz and Weiss, an increase in the rate of interest or an increase in the collateral that entails a shift from region I to region II will result in investors of type a dropping out of the market. However, the less-risky borrowers are more profitable for the bank. Since the bank's rate of return on a loan to entrepreneur i is a concave function of $\tilde{R}_i$, one obtains

$$(6)\quad \rho_a(\gamma) > \rho_b(\gamma) \quad \text{if} \quad C < (1+r)B.$$

Hence, even when there is an excess demand for loans it may not be profitable for a bank to enter the market by raising the rate of interest r, *or* the collateral requirements C. The reason is that only the less-profitable borrowers of type b might find this offer attractive. Therefore, Stiglitz and Weiss argue that credit may be rationed in a competitive equilibrium.

In the following, it will be assumed that investor i will first apply for his preferred contract if he faces a pair $(\gamma_\alpha, \gamma_\beta)$ of credit offers. But, should he be denied credit, he may apply also for the other contract. Therefore, the bank's total profits on, say, contract γ_α depend also upon the fraction of borrowers who are rationed at γ_β. The fraction of firms that receive credit under the terms of γ_j when applying for it will be denoted by λ_j, where $o < \lambda_j \leq 1$.

The tuple $\{(\gamma_\alpha^*, \gamma_\beta^*), (\lambda_\alpha^*, \lambda_\beta^*), \pi^*\}$ is called a *credit market equilibrium* if, when borrowers choose among contracts to maximize expected profits: (*i*) each contract γ_α^* and γ_β^* yields zero profits to the bank; (*ii*) any additional credit offer γ will make no profits; and (*iii*) there is no excess supply of funds.

By requirement (*iii*) of this definition, it is assumed that competition among depositors will lower the deposit interest rate π as long as $L^s(\pi) > L^d(\gamma_\alpha, \gamma_\beta)$. Note that this does not preclude the possibility of an excess demand for loans. *Credit rationing* is said to occur if some entrepreneur i faces a positive probability of being rejected at each contract γ_j^* which maximizes his expected profits and, at the same time, $\Pi_i(\gamma_j^*) > (1+\pi^*)W$.

An equilibrium will not exhibit rationing if entrepreneurs can provide collateral at no cost. Indeed, if collateral requirements cause no loss in welfare, then lenders can avoid any risk associated with a loan by offering contracts that lie on the $C = (1+r)B$ line in Figure 1. The default rate for these loans is zero and, therefore, no adverse selection occurs. Banks can profitably exploit any excess demand for credit by raising r and C simultaneously along $C = (1+r)B$ and rationing disappears in equilibrium. This shows that a credit market equilibrium with $\gamma^* = \gamma_\alpha^* = \gamma_\beta^*$ may always be found on the $C = (1+r)B$ line if $k = 0$. It satisfies $\rho_a(\gamma^*) = \rho_b(\gamma^*) = \pi^*$, and there is no rationing.

This solution to overcome the problem of adverse selection will not be efficient if the use of collateral implies costs. Banks may, however, use contracts with different collateral requirements as a self-selection mechanism. This is possible if the preferences of investors depend systematically upon their type.[9] Entrepreneur i's marginal rate of substitution between r and C at contract γ is given by

$$(7)\quad \sigma_i(\gamma) = -\frac{F_i((1+r)B - C) + k}{[1 - F_i((1+r)B - C)]B}.$$

[9]See, for example, Michael Spence (1973) and Riley (1979).

By second-order stochastic dominance, one must have $F_b(R) > F_a(R)$ for R sufficiently low. Hence, if $F_b(R) - F_a(R)$ does not change sign over $[0, (1+r)B - C]$, entrepreneurs of type a will exhibit a higher marginal rate of substitution at γ than entrepreneurs of type b. This means that they are inclined to accept a higher increase in collateral for a given reduction in interest payments than entrepreneurs of type b. Therefore, the collateral can be used to reveal the riskiness of an entrepreneur's project. In the following, it will be assumed that $F_b(R) - F_a(R)$ is positive over $(0, \bar{R}_a - W)$. This condition ensures that[10] $\sigma_a(\gamma) > \sigma_b(\gamma)$ everywhere in region I of Figure 1, because entrepreneur a will not undertake a project if $(1+r)B - C > \bar{R}_a - W$.

The fact that the indifference curve of investor b is steeper than investor a's at any contract γ enables banks to offer a pair of different, incentive compatible contracts as,[11] for example, $(\gamma_\alpha^*, \gamma_\beta^*)$ in Figure 2. Moreover, it implies that competition between banks will not be restricted to separate variations of either the rate of interest or the collateral. For this reason, credit rationing will turn out not to be viable against competition. In addition to the entrepreneurs' indifference curves a, a' and b, b', Figure 2 depicts the bank's iso-rate of expected return schedule for loans to each type of borrower as α, α' and β, β', respectively. Hence, for example, $\rho_a(\gamma)$ remains constant along α, α'. The bank's indifference curve for loans to type-i investors has a slope of

$$(8) \quad \mu_i(\gamma) = -F_i((1+r)B - C) \,/ [1 - F_i((1+r)B - C)] B$$

at contract γ in (r, C) space. For $k > 0$ it is less steep than investor i's indifference curve at γ. In Figure 2, contract γ_α^* and γ_β^* are depicted so as to satisfy the conditions of the following theorem.[12]

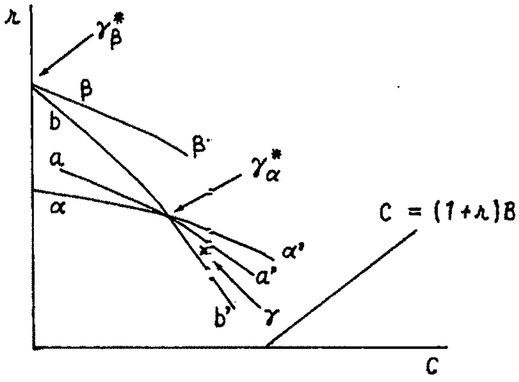


FIGURE 2

THEOREM 1: *Let* $\{(\gamma_\alpha^*, \gamma_\beta^*), (\lambda_\alpha^*, \lambda_\beta^*), \pi^*\}$ *be a credit market equilibrium and let both contracts* γ_α^* *and* γ_β^* *be demanded by entrepreneurs. Then there is no rationing at* γ_α^* *or at* γ_β^*, *and both contracts are incentive compatible. Moreover,* $\pi^* = \rho_a(\gamma_\alpha^*) = \rho_b(\gamma_\beta^*)$. *If* $k > 0$, *then* $C_\alpha^* > C_\beta^* = 0$.

A simple argument establishes that there cannot be rationing at γ_α^* or γ_β^* in Figure 2. A type-a borrower feels rationed at γ_α^* only if $\lambda_\alpha^* < 1$ and $\Pi_a(\gamma_\alpha^*) > (1+\pi^*)W$. However, in such a situation, a competing bank could enter the market and raise r_α^* by a small amount. All borrowers who are denied credit at γ_α^* would apply for this new offer. Since it yields higher returns to the lender than γ_α^*, market entry would be profitable, a contradiction to requirement (*ii*) of an equilibrium. This proves that entrepreneurs of type a are never rationed in equilibrium.

Suppose that some investors of type b are rejected at γ_β^*. Since these borrowers will apply also for γ_α^*, there must be pooling of different entrepreneurs at γ_α^*. Now, consider contract γ in Figure 2 being offered in ad-

[10] Notice that this condition is already implied by (1) in the case of two-outcome projects, i.e., if $\tilde{R}_a$ and $\tilde{R}_b$ are discrete random variables with only two possible realizations. Moreover, the condition would always be satisfied if, instead of mean-preserving spreads, the case of first-order stochastic dominance is considered.

[11] Of course, an identical argument can be developed for n different types of borrowers.

[12] As shown by Rothschild and Stiglitz (1976) and Charles Wilson (1977), a competitive equilibrium may not exist. My Theorem 1, however, may be extended to Riley's "Reactive Equilibrium," or to Wilson's "Anticipatory Equilibrium" (see my 1984a paper).

dition to γ_α^* and γ_β^*. This offer will attract all firms of type a, but only those of type b who are rejected both at γ_β^* and γ_α^*. Therefore, the relative number of type-a to type-b loan applicants will be higher at γ than at γ_α^*. According to (6), entrepreneurs of type a are more profitable for the lender than those of type b. Hence, if γ_α^* yields zero profits, contract γ can be chosen to make a positive profit for the bank, contradicting the definition of equilibrium. This proves that there can be no rationing or pooling if γ_α^* and γ_β^* in Figure 2 constitute an equilibrium. The zero-profit condition, therefore, implies $\pi^* = \rho_a(\gamma_\alpha^*) = \rho_b(\gamma_\beta^*)$.

It remains to show that the set $\{\gamma_\alpha^*, \gamma_\beta^*\}$ is in fact the only possible equilibrium. Among all contracts satisfying $\rho_b(\gamma) = \pi^*$, γ_β^* is the most preferred by type b investors. There is also no contract that could make entrepreneurs of type a better off than γ_α^* without either rendering losses to the bank or attracting high risks from γ_β^*. Therefore, any other configuration of loan offers, which yields zero profits to the lender, is either dominated by $(\gamma_\alpha^*, \gamma_\beta^*)$ or it involves pooling of different investors at some contract. Since, according to the above argument, pooling is never viable against competition, any equilibrium must satisfy the conditions of Theorem 1.

Theorem 1 explains the widespread use of collateral requirements in debt contracts. Since collateral in general is costly, its use is inefficient under perfect information. However, collateral requirements can be explained as a response to imperfect information. They may serve to reveal information about the default risk of loan applicants. High-risk borrowers can be identified because they prefer loan contracts with lower collateral and a higher interest rate. In an equilibrium, the level of collateralization is negatively related to the riskiness of the borrower's investment project.

III. Conclusions

Problems of adverse selection cast serious doubts on the appropriateness of the Walrasian equilibrium concept in markets with imperfect information. If the average quality of supplied goods is positively related to the market price, then buyers have an incentive to fix the price above the market-clearing level. In this situation, sellers become rationed because buyers prefer to pay a higher price in exchange for higher quality. As another response to adverse selection, a signalling convention may emerge. Sellers of high-quality goods invest in observable characteristics to distinguish their products from those of lower quality. In this case, the simple Walrasian auction market is replaced by a more complicated market for contracts. Typically, the price at which goods are exchanged is only one among several other characteristics of such contracts. This paper has investigated the question how the market reacts if both signalling and rationing are a priori feasible alternatives. In the framework of credit markets under imperfect information, it has been established that no borrower will be denied credit if banks use the collateral requirements of their loan contracts as a signalling mechanism. Signalling mechanisms eliminate demand rationing in competitive equilibrium. This suggests that theories which attempt to explain rationing by adverse selection effects also have to provide an explanation why screening devices cannot be adopted.

In this paper, the applicability of self-selection mechanisms has been established under assumptions which are stronger than those of Stiglitz and Weiss. First, a signalling equilibrium in the credit market necessitates a monotone relationship between the riskiness and the preferences of different borrowers. It has been shown that this condition is satisfied under an additional assumption on the probability distribution of returns on investment projects. Second, low-risk entrepreneurs have been assumed to be able to raise a sufficient amount of collateral to distinguish themselves from high risk ones. This is important because, obviously, perfect sorting in a credit market equilibrium may be impossible if some low-risk firms face a binding constraint on the amount of collateral they can provide.[13] Thus, only partial screen-

[13] The results of Section II remain valid, however, if the maximum amount of collateral is determined by the assumption that the marginal costs of a higher collateral are increasing and tend to infinity at some level $\bar{C}$.

ing may be possible and adverse selection could still arise, if the necessary conditions for market signalling are not fulfilled.

REFERENCES

Bester, Helmut, (1984a) "Screening versus Rationing in Credit Markets with Imperfect Information," Discussion Paper No. 136, University of Bonn, February 1984.

______, (1984b) "The Level of Investment in Credit Markets with Imperfect Information," Discussion Paper No. 150, University of Bonn, August 1984.

Jaffee, Dwight M. and Russell, Thomas, "Imperfect Information, Uncertainty, and Credit Rationing," *Quarterly Journal of Economics*, November 1976, *90*, 651–66.

Milde, Hellmuth and Riley, John G., "Signalling in Credit Markets," Working Paper No. 334, University of California-Los Angeles, July 1984.

Riley, John G., "Informational Equilibrium," *Econometrica*, March 1979, *47*, 331–59.

Rothschild, Michael and Stiglitz, Joseph, "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *Quarterly Journal of Economics*, November 1976, *90*, 630–49.

______ **and** ______, "Increasing Risk: I, A Definition," *Journal of Economic Theory*, September 1970, *2*, 225–43.

Spence, Michael, *Market Signalling: Information Transfer in Hiring and Related Processes*, Cambridge: Harvard University Press, 1973.

Stiglitz, Joseph and Weiss, Andrew, "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, *71*, 393–410.

Wette, Hildegard C., "Collateral in Credit Rationing in Markets with Imperfect Information: Note," *American Economic Review*, June 1983, *73*, 442–45.

Wilson, Charles, "A Model of Insurance Markets with Incomplete Information," *Journal of Economic Theory*, December 1977, *16*, 167–207.

The Conformity of Wage-Indexation Models with "Stylized Facts"

By GARY FETHKE*

In a recent paper, Bennett McCallum (1982) lists what he considers to be prominent empirical regularities or "stylized facts" of aggregate economies. In particular, he notes that "...output and employment magnitudes are strongly related to contemporaneous money stock surprises, [but that] ...output and employment magnitudes are not strongly and positively related to contemporaneous price level surprises" (p. 4). These facts have prompted McCallum and others to develop models of the economy where prices as well as wages are predetermined. The main feature of such models is their abandonment of aggregate-supply formulations where price level disturbances provide a channel for the real effects of money.

The purpose of this paper is to demonstrate the consistency of familiar Gray and Fischer wage-indexing models and their implied aggregate-supply relationship with the stylized facts (see JoAnna Gray, 1976; Stanley Fischer, 1977). In a model where the nominal wage is indexed to the price level, the efficient use of the information conveyed by the price level imposes qualitative restrictions on the covariance matrix of disturbances. First, the correlation between the price level and innovations in the deviation of actual output from the full-information output level will be zero. Second, because the money supply contains information about real disturbances that is not conveyed by the price level, the correlation between money supply innovations and innovations in the deviation of output from the full-information level will be positive. Third, the regression of innovations in *actual* output on the price level will provide an estimate of the optimal degree of indexation. Evidence that is generally consistent with these properties of wage-indexation models is found in quarterly data for the five largest OECD countries.

Consider the familiar aggregate formulation:[1]

$$(1)\quad y_t = c_1(p_t - p_{t|t-1}) + (1+c)\varepsilon_t \qquad c_1, c > 0,$$

$$(2)\quad y_t = -ar_t + a(p_{t+1|t-1} - p_t) + u_t \qquad a > 0,$$

$$(3)\quad m_t - p_t = b_1 y_t - b_2 r_t + v_t \qquad b_1, b_2 > 0.$$

Here, y_t is the logarithm of real output, r_t is the nominal rate of interest, m_t is the logarithm of the money stock, p_t is the logarithm of the price level, and $p_{t|t-1}$ and $p_{t+1|t-1}$ denote the mathematical expectations of p_t and p_{t+1} formed at period $t-1$. The information set at $t-1$ includes knowledge of the model's structure and values of all variables dated $t-1$ and earlier; it does not include information about t-period variables. The disturbance terms ε_t, u_t, and v_t are uncorrelated white noise processes, with variances σ_ε^2, σ_u^2, and σ_m^2. For convenience, all constants are suppressed.

The aggregate-supply schedule, equation (1), presumes that wages are established at the end of period $t-1$ so as to equate the expected demand and supply of labor. Employment is determined *ex post* by the labor demand schedule. Substitution of the prede-

*University of Iowa, Iowa City, IA 52242. I thank Andrew Policano and John Kennan for helpful comments.

[1] Except for the rationalization of the aggregate-supply schedule, this model is similar to that of Thomas Sargent and Neil Wallace (1975). For a complete development of the aggregate-supply schedule for both the cases of a predetermined and indexed nominal wage, see my article with R. Jackman (1984).

termined nominal wage into the labor demand schedule and then employment (labor demand) into the production function yields equation (1). For simplicity, labor supply is assumed fixed which implies that the full-information level of output is determined as $y_t^* = \varepsilon_t$.

In equation (2), real income depends on the real rate of interest with investors forming their expectations of the price level in period $t+1$ based on information available at time $t-1$. Equation (3) is the portfolio-balance equation.

To complete the model, either the money supply or the interest rate can be selected as the policy instrument. The choice of policy instrument, however, will affect the information content of the price level (see my article with R. Jackman). Under a money supply rule, disturbances in the money market will influence both the price level and the level of output through their effect on the interest rate. Alternatively, under an interest rate rule, output and the price level are isolated from random disturbances in the money market. Since one concern here is with the relationship between output and money supply innovations for the case where the money supply conveys information about productivity disturbances, the interest rate is selected as the policy instrument. The case of a money supply rule is presented in the Appendix.

To achieve a determinate price level under an interest rate rule, it is assumed, following an argument provided by McCallum (1981), that the monetary authority selects the value of the interest rate such that

$$(4) \quad r_t = b_2^{-1}\left(b_1 y_{t|t-1} + p_{t|t-1} - m_t^*\right),$$

where m_t^* is an exogenously determined target level of the money stock. Thus, the interest rate is set at $t-1$ to equate the expected demand for money with the target money stock. For simplicity, let $m_t^* = 0$.

It is convenient to first solve the model based on information available at $t-1$ and then to introduce the current-period information that results from indexation of nominal wages to the price level. In this formulation, $p_{t|t-1} = p_{t+1|t-1} = 0$.[2] The rational expectations solution for the price level, the money supply, and the deviation of output from the full-information level are

$$(5) \quad p_t = \left[u_t - (1+c)\varepsilon_t\right] \div (a + c_1),$$

$$(6) \quad m_t = \left[(1+c)(ab_1 - 1)\varepsilon_t + (1 + b_1 c_1)u_t + (a + c_1)v_t\right] \div (a + c_1),$$

$$(7) \quad y_t - y_t^* = \left[c_1 u_t + (ca - c_1)\varepsilon_t\right] \div (a + c_1).$$

In the standard models of wage indexation, the nominal wage is adjusted to reflect the information contained in p_t. No other current-period aggregate variable is observed.[3] In equation (1), the slope of the supply schedule c_1 is initially treated as a fixed parameter. If the nominal wage is indexed to the current price level, instead of being predetermined, then the slope of the supply schedule will depend on other parameters of the model. Define $c_1 \equiv c(1-\gamma)$ where γ is the indexation parameter. Indexation of the wage rate to the price level involves using the information conveyed by the price level

[2] Substitute equation (4) into equation (2) and then solve equations (1) and (2) for the price level:

$$(a + c_1)p_t = (c_1 - c/b_2)p_{t|t-1} + ap_{t+1|t-1} - (ab_1/b_2)y_{t|t-1} - (1+c)\varepsilon_t + u_t.$$

Forward this expression j periods and then take expectations of p_{t+j} conditional on period $t-1$ information; this yields $p_{t+j|t-1} = (b_2/(1+b_2))p_{t+j+1|t-1}$. Set $j=1$ and solve forward to obtain $p_{t+1|t-1} = (b_2/(1+b_2))^{N-1}p_{t+N|t-1}$. If $p_{t+N|t-1}$ is finite, then $p_{t+1|t-1} = 0$ for large N. To obtain the result that $p_{t|t-1} = 0$, substitute $p_{t+1|t-1} = 0$ into equation (2) and then form expectations of equations (1)–(3) based on period $t-1$ information, with r_t given by equation (4). Solving the resulting equations gives $p_{t|t-1} = 0$.

[3] It is easy to demonstrate that observation of p_t and m_t under the interest rate regime will reveal the productivity disturbance, ε_t, and thus reveal the natural rate of output. It is more difficult to understand, since the wages can be indexed to the price level, why wages are not also indexed to the money supply; see Benjamin Eden (1983).

to form the minimum mean square, "optimal," estimate of output. The degree of indexation is chosen to equate the conditional expectation of actual output to the conditional expectation of full-information output; that is, the indexation parameter, γ, is selected such that $E[(y_t - y_t^*)|p_t] = 0$. The resulting value for the optimal degree of indexation is given by

$$(8) \quad \gamma = 1 - \frac{(1+c)a\sigma_\varepsilon^2}{\sigma_u^2 + (1+c)\sigma_\varepsilon^2}, \qquad \text{with} \quad a \leq 1 - \gamma \leq 1.$$

The optimal use of the information contained in the price level imposes restrictions on the covariances between innovations in the deviation of output from the full-information level, innovations in the money stock, and disturbances in the price level. Since the degree of indexation is found by setting the indexation parameter and thereby the slope of the supply schedule such that $E[(y_t - y_t^*)|p_t] = 0$, optimal indexation implies that innovations in output deviations from the full-information level are not correlated with the current price level.

On the basis of information available at $t-1$, the covariance between innovations in actual output and innovations in the price level (computed from equations (5) and (7)) is

$$(9) \quad \mathrm{Cov}(y_t, p_t) = \frac{c_1\sigma_u^2 - a(1+c)^2\sigma_\varepsilon^2}{(a+c_1)^2} \lesseqgtr 0.$$

When the information contained in p_t is efficiently used to index wages, the covariance between y_t and p_t is negative. Using equations (8), (9), and $c_1 \equiv c(1-\gamma)$,

$$(10) \quad \mathrm{Cov}(y_t, p_t) = \frac{-(1+c)\sigma_\varepsilon^2(\sigma_u^2 + (1+c)\sigma_\varepsilon^2)}{a(\sigma_u^2 + (1+c)^2\sigma_\varepsilon^2)} < 0.$$

In fact, optimal indexation of wages to the price level imposes an identification restriction in the regression of actual output innovations on the price level. With wages optimally indexed,

$$(11) \quad E(y_t|p_t) = \left[\frac{-a(1+c)\sigma_\varepsilon^2}{\sigma_u^2 + (1+c)\sigma_\varepsilon^2}\right]p_t.$$

Here, one plus the population regression coefficient equals the optimal degree of wage indexation, γ (see equation (8)). If the aggregate-demand function is stable, $\sigma_u^2 = 0$, then productivity shocks ε_t will shift aggregate supply and yield a correlation of minus one between the price level and output innovations. If the aggregate-supply function is stable, $\sigma_\varepsilon^2 = 0$, then optimal indexation implies a zero correlation between price and output innovations. The general case requires a negative correlation.

When wages are optimally indexed to the price level, innovations in output deviations will be correlated with innovations in the money stock. Under the pure interest rate rule (equation (4)), money supply shocks v_t do not affect either output or the price level; however, the current money stock disturbance does contain information about $y_t - y_t^*$ that is not contained in the price level.[4] Specifically, the covariance between innovations in the money supply and innovations in the deviations of output from the full-information level is determined from equations (6), (7), and (8), and is given by

$$(12) \quad \mathrm{Cov}(y_t - y_t^*, m_t) = \frac{b_1c(1+c)\sigma_\varepsilon^2\sigma_u^2}{\sigma_u^2 + (1+c)^2\sigma_\varepsilon^2} > 0.$$

This configuration of optimal wage indexation and a noncontingent interest rate rule provides one example of an economy where

[4]Under a money supply rule, in contrast to the interest rate rule, disturbances in the money market will affect output; see the Appendix. The choice between policy rules depends on which rule, when combined with optimal wage indexation, will provide the most information about real disturbances. If, for example, there are substantial disturbances in the money market relative to the goods market, an interest rate rule that isolates the labor and goods markets from disturbances in the money market will be preferred to the money-supply rule. These issues are addressed by myself and Jackman.

the money supply has no direct impact on output, but where it is possible to identify a positive correlation between contemporaneous surprises in output deviations and surprises in the money supply. Furthermore, even though price disturbances have a direct effect on output through the aggregate-supply function, equation (1), the correlation between the price level and output surprises, measured as deviation from the full-information level, is nonexistent. These restrictions on the covariances of innovations result from the efficient use of price level information and not from the particular mechanism chosen to incorporate that information. If the interest rate or the money supply were optimally contingent on the price level, the same level of efficiency as that achieved under optimal wage indexation would result (see my article with Jackman).

Table 1 presents estimates of the contemporaneous correlations between output and price level innovations for the five largest OECD countries, using quarterly data from 1955:I to 1981:IV. The first column presents the correlation between the first differences of the logarithms of output and the price level; this would be the correct specification if the logarithms of output and the price level each followed random walk processes. The second column represents the contemporaneous correlation of the innovations from a three-variable vector autoregression of order two, where the variables are the logarithms of output, the price level, and the money supply. The vector autoregression also includes a quadratic trend and seasonal dummy variables.

As predicted by a model in which current price level information is used to forecast current output, the correlations between the price disturbances and the disturbances in actual output are typically negative: for all five countries and for both measures of the contemporaneous correlation between output and price level disturbances, the relationship is estimated to be negative.[5] The estimate of the optimal degree of indexation, measured by adding one to the estimate of the regression coefficient in a regression of output innovations on the price disturbances, is given by the final column of Table 1. The result for the United States, which indicates

TABLE 1—CONTEMPORANEOUS CORRELATIONS BETWEEN INNOVATIONS IN OUTPUT AND THE PRICE LEVEL FOR FIVE OECD COUNTRIES, QUARTERLY DATA, 1955:I–1981:IV[a]

Country	$\text{Corr}(\Delta y, \Delta p)$	$\text{Corr}(\tilde{y}, \tilde{p})$	$\hat{\gamma}$
Canada	$-.249$	$-.174$	.766
Germany[b]	$-.134$	$-.057$	.825
Japan	$-.086$	$-.020$	.963
United Kingdom	$-.295$	$-.137$	.848
United States	$-.342$	$-.414$	$-.051$

[a]The price and output series are the logarithms of real *GDP* (or *GNP*) and the logarithm of the *GDP* (or *GNP*) deflator for all countries except Germany, where output is measured by the logarithm of the index of industrial production and the price level by the logarithm of the producer price index. The data are from the *National Accounts Statistics*, *National Accounts for OECD Countries*, various issues. The following notation is used: Δy and Δp measure the first difference of the logarithms of output and the price level; $\tilde{y}$ and $\tilde{p}$ are the innovations in output and the price level that derive from a three-variable vector autoregression of order two, where the variables are the logarithms of output, the price level, and the money supply; finally, $\hat{\gamma}$ measures the implied degree of indexation formed by adding one to the regression coefficient in the regression of output innovations on price level innovations.

[b]Using *GNP* data rather than industrial production for the period 1960:I–1981:IV, $\text{Corr}(\Delta y, \Delta p) = -.349$ and $\text{Corr}(\tilde{y}, \tilde{p}) = -.188$. The implied degree of indexation for this period is $\hat{\gamma} = .506$.

[5]Charles Nelson (1979) and Christopher Sims (1980) estimate negative covariances between innovations in output and the price level for the postwar U.S. economy. The evidence, across economies, using monthly data on industrial production and the wholesale price index is mixed; see Patrick Geary and John Kennan (1982). Kennan pointed out to me that the negative correlation between output and price level innovations may be partially the result of the measurement error that is introduced when real *GDP* is determined by dividing nominal *GDP* by the *GDP* deflator. In logarithms, let $\bar{x} \equiv \bar{p} + \bar{y}$ where $\bar{x}$, $\bar{p}$ and $\bar{y}$ are the true values of nominal *GNP*, the price level and real *GNP*. If $x = \bar{x} + \eta$ and $p = \bar{p} + \varepsilon$, where $\text{Var}(\eta)$ and $\text{Var}(\varepsilon) < \infty$ and $\text{Cov}(\eta\bar{p}) = \text{Cov}(\varepsilon\bar{p}) = \text{Cov}(\varepsilon\bar{x}) = 0$, then $\text{Cov}(py) = \text{Cov}(\bar{p}\bar{y}) + \text{Cov}(\varepsilon\eta) - \text{Var}(\varepsilon)$. The effect of measurement cannot be determined. Given the way that nominal *GNP* and the *GNP* deflator are measured, however, there is no reason to suspect any particular covariation between ε and η.

essentially no indexation for the aggregate economy, contrasts with that of the other OECD economies, where the degree of indexation implies a rigid real wage.

In summary, there is support in these data for the prediction that the optimal use of price level information in an economy subject to both demand and supply shocks will lead to a nonpositive correlation between innovations in actual output and the price level. This prediction is compatible with an economy where the price level is flexible (i.e., not predetermined) and where the aggregate-supply schedule exhibits a positive slope based on a transactions-cost model for determining the nominal wage. The slope of the aggregate-supply schedule is determined endogenously and depends, through wage indexation, on the variances of demand and supply disturbances.

Appendix: The Case of a Money Supply Rule

For convenience, represent equation (3) as $m_t - p_t = b_1 y_t - b_2 r_t$, where m_t is a white noise process that is uncorrelated with u_t and ε_t. The rational expectations solutions for the price level, the interest rate, and the deviation of output from the full-information level are given by

$$\text{(A1)}\quad p_t = [b_2 u_t + a m_t - (1+c)(ab_1+b_2)\varepsilon_t] \div [a(1+b_2) + (ab_1+b_2)c_1],$$

$$\text{(A2)}\quad r_t = [(1+b_1c_1)u_t - (a+c_1)m_t - (1+c)(1-ab_1)\varepsilon_t] \div [a(1+b_2) + (ab_1+b_2)c_1],$$

$$\text{(A3)}\quad y_t - y_t^* = [(a(1+b_2)c - (ab_1+b_2)c_1)\varepsilon_t + (b_2u_t + am_t)c_1] \div [a(1+b_2) + (ab_1+b_2)c_1],$$

where, again, $p_{t|t-1} = p_{t+1|t-1} = 0$.

With $c_1 \equiv c(1-\gamma_m)$ and γ_m selected such that $E[(y_t - y_t^*)|p_t] = 0$,

$$\text{(A4)}\quad \gamma_m = 1 - \frac{a(1+b_2)(ab_1+b_2)(1+c)\sigma_\varepsilon^2}{b_2^2\sigma_u^2 + a^2\sigma_m^2 + (1+c)(ab_1+b_2)^2\sigma_\varepsilon^2}.$$

When wages are optimally indexed, $\mathrm{Cov}(y_t - y_t^*, p_t) = 0$, however,

$$\text{(A5)}\quad \mathrm{Cov}(y_t - y_t^*, m_t) = \frac{ca(ab_1+b_2)(1+c)\sigma_m^2\sigma_\varepsilon^2}{b_2^2\sigma_u^2 + a^2\sigma_m^2 + (1+c)^2(ab_1+b_2)^2\sigma_\varepsilon^2} > 0.$$

The regression of actual output on the price level yields

$$\text{(A6)}\quad E[y_t|p_t] = \left[\frac{-a(1+b_2)(ab_1+b_2)(1+c)\sigma_\varepsilon^2}{b_2^2\sigma_u^2 + a^2\sigma_m^2 + (1+c)(ab_1+b_2)^2\sigma_\varepsilon^2}\right]p_t.$$

One plus the regression coefficient equals γ_m. While the indexation parameters differ, the qualitative results regarding the covariances of innovations are the same as those for the interest rate rule.

REFERENCES

Eden, Benjamin, "Competitive Price Setting, Price Flexibility and Linkage to the Money Supply," in Karl Brunner and Alan Meltzer, eds., *Variability in Employment, Prices, and Money*, Carnegie-Rochester Conference Series on Public Policy, Vol. 19, Amsterdam: North-Holland, 1983, 253–300.

Fethke, Gary and Jackman, R., "Optimal Monetary Policy, Endogenous Supply and Rational Expectations," *Journal of Monetary Economics*, March 1984, *13*, 211–24.

Fischer, Stanley, "Wage Indexation and Macroeconomic Stability," in Karl Brunner and Alan Meltzer, eds., *Stabilization of the*

Domestic and International Economy, Carnegie-Rochester Conference Series on Public Policy, Vol. 5, *Journal of Monetary Economics*, Suppl. 1977, 107–47.

Geary, Patrick T. and Kennan, John, "The Employment-Real Wage Relationship: An International Study," *Journal of Political Economy*, August 1982, *90*, 854–71.

Gray, JoAnna, "Wage Indexation: A Macroeconomic Approach," *Journal of Monetary Economics*, April 1976, *2*, 221–35.

McCallum, Bennett T., "Price Level Determinancy with an Interest Rate Policy Rule and Rational Expectations," *Journal of Monetary Economics*, November 1981, *8*, 319–329.

______, "Macroeconomics After a Decade of Rational Expectations: Some Critical Issues," *Economic Review*, Federal Reserve Bank of Richmond, November/December 1982, *68*, 3–12.

Nelson, Charles R., "Recursive Structure in U.S. Income, Prices, and Output," *Journal of Political Economy*, December 1979, *87*, 1307–27.

Sargent, Thomas J. and Wallace, Neil, "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rules," *Journal of Political Economy*, April 1975, *83*, 241–54.

Sims, Christopher A., "Comparison of Interwar and Postwar Business Cycles: Monetarism Reconsidered," *American Economic Review Proceedings*, May 1980, *70*, 250–57.

Auctions with Contingent Payments

By ROBERT G. HANSEN*

There now exists a host of results concerning the revenue performance of various auction methods. This note delves deeper into auction markets by examining the effects on sellers' revenue of certain noncash means of payment. The basic result—that bidding mediums which include some contingent pricing feature generally yield the seller more revenue than do cash bids—is intriguing by itself and also points out the limitations of received theory.

Since what follows builds on the independent-preferences framework, it is useful to first note the assumptions of that model and the major results pertaining to it. To model an auction in the independent-preferences tradition, one assumes that bidders have reservation values, V_i, that are known only privately and that can be depicted as being drawn independently from some distribution $F(V)$. The most important result for this model is the revenue equivalence theorem: as given in John Riley and William Samuelson (1981), any auction involving risk-neutral bidders for which the following four conditions hold:

(a) a buyer can make any bid above some minimum "reserve" price,

(b) the buyer making the highest bid is awarded the object,

(c) the auction rules are anonymous, and

(d) there is a common equilibrium bidding strategy in which each buyer makes a bid b_i, which is a strictly increasing function of his reservation value V_i, yields an expected revenue of

$$(1)\quad E(revenue) = \int_{V^*}^{V} (VF'(V)+F(V)-1)F(V)^{n-1}dV,$$

where n = number of bidders and V^* = reservation value below which it is unprofitable to submit a bid.[1]

The limitations of this result can be pointed out by considering two common noncash bidding methods, stock bidding in the market for corporate control, and profit-share bidding in oil lease auctions. Under independent-preferences assumptions (and the additional condition that some variable correlated with the V_i becomes observable *ex post*), both of these auction methods yield expected revenue strictly greater than that given by (1).[2] The examples are quite simple yet it should be clear that they imply a general result on the dominance of contingent-payment bidding mediums.

I. Stock Bidding

Consider an auction involving one "target" firm and n potential "acquiring" firms. Each acquiring firm has value of itself equal to X (X will later be allowed to vary); the target has value to acquiring firm A_i of V_i. To meet independent-preferences assumptions, assume that the V_i can be represented as independent draws from the distribution $F(V)$, that V_i is known only to A_i, and that X is publicly known.

Suppose then that the target firm is sold by an open (progressive), cash-bid auction with zero reserve price. From standard auction theory, the high bid b^* for this auction will be

$$(2)\qquad b^* = V_2,$$

where V_2 equals the second highest of the V_i.

*Amos Tuck School of Business Administration, Dartmouth College, Hanover NH 03755. I thank Richard Bower, Dennis Logue, John Riley, and two referees for helpful comments. Financial support was received from the Tuck Associates.

[1] See Riley and Samuelson (pp. 382–83).

[2] The findings here do not technically contradict Riley and Samuelson's proposition because their (implicit) assumption is that V_i is unobservable even *ex post*. Also, for the auction rules considered here, property (c) concerning anonymity does not hold: the same bid can entail different payments for different bidders.

Furthermore, the auction will be won by the firm that values the target highest, and the seller's expected revenue equals the expectation of the second-highest value.

Suppose instead that the target requires open bidding of stock offers, where a stock offer is defined as the percentage of the merged entity that the target will own after trade. The highest-percentage bid naturally decides the winner.

Then it is clear that a dominant bidding strategy is to bid up to

$$(3) \qquad p(V_i)=(V_i/(X+V_i)).$$

The firm with the highest V_i still wins this auction, and the high bid p^* will be

$$p^*=p(V_2)=(V_2/(X+V_2)),$$

where V_2 is again the second-highest value. Interestingly, the value of p^* to the target is given as

$$\begin{aligned} \textit{value to target} &= \frac{V_2}{X+V_2}(X+V_1)>V_2 \\ &= \textit{high cash bid}, \end{aligned}$$

where V_1 is the highest value. For every possible set of V_i, the open stock auction therefore yields more revenue than the open cash auction (the revenue for which equals revenue from other common auction rules, for example, cash-sealed high bid).[3]

Of course, the assumption that all acquiring firms have the same value X is not very realistic. However, we can allow acquiring firm i to have value X_i—as long as this is known publicly[4]—and a stock auction will still yield more revenue than a cash auction (which is unaffected by acquiring-firm values). The auction procedure now becomes more complex, but, at the same time, more interesting.

For this new model, the bidding variable for an open auction is still p_i, the percentage of ownership that firm i offers. Now, however, the target cannot simply choose the highest p-offer, for firms with low X_i can make high p-offers. Suppose that the target announces that p-offers will be evaluated according to a function Z, where

$$(4) \qquad Z(p_i)=(p_iX_i/(1-p_i)).$$

The firm offering the highest Z will win and pay its associated p-offer.

Faced with this evaluation rule, acquiring firms have a dominant strategy of bidding up to

$$(5) \qquad p_i^*=(V_i/(X_i+V_i)).$$

There would be no reason to stop bidding at a lower p and forego possible gains; likewise, there would be no reason to bid more than p_i^* since upon winning that would entail paying more than value.

Notice that (5) implies each firm offers a Z up to

$$(6) \qquad Z_i^*=\left(p_i^*X_i/\left(1-p_i^*\right)\right)=V_i$$

This is, of course, why the target's rule (4) is a sensible one. Furthermore, it then follows that the firm with the highest V_i (call this firm W) will win the auction and will have to offer a p_w such that

$$(7) \qquad (p_wX_w/(1-p_w))=V_2$$

since the second highest of the Z_i will equal V_2. Equation (7) implies

$$p_w=(V_2/(X_w+V_2)),$$

so that the value of the winning p-offer to the target is given by

$$\textit{value of } p_w \textit{ to target} = p_w(X_w+V_1)>V_2.$$

The stock auction still dominates a cash auction.

II. Profit-Share Bidding in Oil Lease Auctions

Here, I must resort to some specific assumptions so that equilibrium does not degenerate to one where everybody bids a

[3] Notice that V^* does not differ between the cash and stock auctions; in both cases $V^*=0$. Also, expected revenue is still equal across open and sealed-bid auctions using stock bids.

[4] My 1984 paper discusses cases where X_i is only known privately.

profit share of 100 percent.[5] Specifically, assume that the lease up for sale either has no oil, or it has oil worth V. All of n potential bidders agree there is a $(1-p)$ chance of no oil and a p chance of "striking" V.[6] Unfortunately for the bidders, the cost to firm i is C_i to find out the state of nature, and sellers share only in profits. To meet the independent preferences assumption, assume that the C_i can be represented as independent draws from a distribution $F(C)$. Firm i then calculates expected value of the lease as $pV - C_i$.

In an open cash auction with zero reserve price, the high bid b^* will be

$$b^* = pV - C_2, \tag{8}$$

where C_2 is the second lowest of the C_i. The auction will, as usual, be won by the firm having the highest value (lowest C_i); expected revenue for the seller is the expected value of b^*.

With open profit-share bidding where the highest share wins, firm i will bid up to[7]

$$q(C_i) = ((pV - C_i)/(pV - pC_i)). \tag{9}$$

The high bid q^* will be

$$q^* = ((pV - C_2)/(pV - pC_2)),$$

and the firm with the lowest C_i will win. Expected value of q^* to the seller is

$$\textit{value of } q^* = ((pV - C_2)$$
$$/(pV - pC_2))(p)(V - C_1) > pV - C_2 = b^*$$

As for stock offers in the merger market, profit-share bidding increases expected revenue.[8]

III. Conclusions

A general principle can be drawn from the results for these two examples. The source of the gain from using contingent-payment bidding methods is that these methods allow the seller to capture some portion of the difference between the two highest reservation values, $V_1 - V_2$. Cash auctions, of course, yield expected revenue equal only to V_2. With contingent payments, the "package" offered by the winning bidder (for instance, the profit share offered) is such that, if applied to the second-highest bidder, would yield that bidder a zero profit upon winning and would yield the seller revenue of V_2. The winning package will not, however, be applied to the second highest but to the first highest; it is this twist that allows the seller to capture some of $V_1 - V_2$.

It should also be pointed out that independent preferences do not capture all the aspects of any auction completely; however, to the extent that values do differ across bidders (for example, even in the generalized independent-preferences/mineral rights model of Paul Milgrom and Robert Weber, 1982), the effects of contingent bidding rules should continue to hold.

REFERENCES

Hansen, Robert G., "Informational Asymmetry and the Means of Payment in Auctions," mimeo., Amos Tuck School, Dartmouth College, 1984.

Mead, Walter J., Moseidjord, Asbjorn and Muraoka, Dennis D., "Alternative Bid Variables as Instruments of OCS Leasing Policy," *Contemporary Policy Issues*, March 1984, 5, 30–44.

[5] Douglas Reece (1979) and Walter Mead et al. (1984) have previously considered profit-share bidding and made similar assumptions (except for the informational assumptions; see fn. 6). Mead et al. cannot conclude that profit-share bidding unambiguously increases revenue; Reece finds an unambiguous increase but does not give an explanation such as that given here.

[6] Notice I am not employing the usual "mineral rights" assumptions of independent estimates on oil potential.

[7] The firm loses C_i without a strike and gets $(1-q)(V - C_i)$ with a strike.

[8] It should be noted that all contingent-pricing rules will involve incentive problems that could reduce the seller's expected revenue. For this reason, royalty-rate bidding in oil lease auctions has not been presented as another example.

Milgrom, Paul R. and Weber, Robert J., "A Theory of Auctions and Competitive Bidding," *Econometrica*, September 1982, *50*, 1089–1122.

Myerson, Roger B., "Optimal Auction Design," *Mathematics of Operations Research*, 1981, *6*, 58–73.

Reece, Douglas K., "An Analysis of Alternative Bidding Systems for Leasing Offshore Oil," *Bell Journal of Economics*, Autumn 1979, *10*, 659–70.

Riley, John G., and Samuelson, William, "Optimal Auctions," *American Economic Review*, June 1981, *71*, 381–92.

Inconsistent Equilibrium Constructs: The Evenly Rotating Economy of Mises and Rothbard

By TYLER COWEN AND RICHARD FINK*

The growing disillusionment with the Arrow-Hahn-Debreu model of general equilibrium has led to an examination of alternative constructions. Among the alternatives that have recently been studied are temporary equilibria, rational expectations equilibria, and the evenly rotating economy (*ERE*).[1] This note will focus on the *ERE*. We will argue that, whatever problems equilibrium constructs may have in general, these problems are compounded by serious inconsistencies in both the nature of the *ERE* and its suggested uses.

We focus on four particular features of the *ERE*. First, it is the result of a convergence process initiated by a freeze of tastes, technology, and resources (Murray Rothbard, 1962, p. 273; Mises does not specify the conditions required for convergence); second, the events of a single market "day" continually repeat themselves (Ludwig Mises, 1949, p. 247; Rothbard, p. 273); third, the *ERE* does not contain the Arrow-Hahn-Debreu restrictions of perfect futures markets and approximate convexities; fourth, the *ERE* contains money and money prices (Mises, pp. 416–17, 538; Rothbard, passim).

Both Mises and Rothbard propose the following uses of the *ERE*:

1) *The ERE can be used to explain or "predict" the direction of change* (Mises, pp. 244–45; Rothbard, pp. 275–76). Since the market economy is supposedly always moving towards the *ERE*, a clear notion of the relevant end-state will tell us what state of affairs the market is tending to establish (but will never actually reach).[2]

2) *The ERE is an analytical building block or stepping stone towards analyzing complex phenomena in a world of change* (Rothbard, p. 276).

3) *The ERE is a starting point for an analysis of the effects of particular changes.* If we wish to analyze the effects of an exogenous shock upon the economic system, we start with a state where there is no change and then allow the new change to work its way through the system until the economy settles into equilibrium once again. Thus, we can isolate the effects of this change (Mises, p. 248).

4) *The ERE is used as a foil.* Since the *ERE* is *so* unrealistic, it is supposed to shed light upon the real world by method of contrast. For instance, examining a world in which there is no uncertainty and, hence, no entrepreneurship, may help shed light on the nature of entrepreneurship in the real world (Mises, p. 248).

I. The *ERE*: A Critique

Using the *ERE* to explain the direction of change in a market rests upon the notion that there is a "tendency" for a market economy to approach equilibrium. "If our *data* —values, technology, and resources—remained constant, the economy would move

*Departments of Economics, George Mason University, Fairfax, VA 22030, and Harvard University, Cambridge, MA 02138, respectively. We thank the Center for the Study of Market Processes for support on this project.

[1] Not only is the *ERE* frequently used in modern "Austrian" literature (see Roger Garrison, 1978, for an example) but is also gaining numerous other adherents. In a recent talk at Harvard University ("Principles of Monetary and Fiscal Policy," 1984) Robert Lucas expressed his admiration for Mises' *ERE* and noted its close similarity with much of his own work with overlapping generations models. However, since Mises and Rothbard contain the primary statement and defense of the *ERE*, we shall focus our attention on their writings. Despite our critical tone, we wish to note our appreciation of the elements of market process and order analysis in Rothbard and Mises.

[2] Mises attaches the label "final state of rest" (p. 245) to this use of the *ERE*.

toward the final equilibrium position and remain there" (Rothbard, p. 275). Although Rothbard later claims that verbal analysis is capable of explaining the convergence path (p. 278), no such explanation is forthcoming. In fact, there is no reason to believe that the economic system will settle into an equilibrium state, for as the analysis of Gerard Debreu (1959) implies, sequential transactions are *not* consistent with the notion of an intertemporal general equilibrium.[3] Instead, all transactions across time must already be perfectly coordinated on the first day of trading. Any attempt to do away with this stipulation is bound to encounter all of the problems that have plagued mathematical general equilibrium convergence theory in the past. Unless the equilibrium of the system is preordained (as it is in Debreu), it is nearly impossible for the learning processes of actors to be so highly proficient and adaptable as to allow for an actual convergence to equilibrium.[4]

Among the specific problems preventing perfect plan coordination are income effects, disequilibrium production and consumption, nonconvexities, strategic behavior (resulting from externalities), and the potentially false price signals generated by the process of convergence. Mathematical general equilibrium theory usually places serious (and unrealistic) restrictions on these problems in order to generate a process of convergence. Even then, it is still often impossible to demonstrate convergence (see Franklin Fisher, 1976). How can the *ERE* explain convergence without such restrictions?[5]

However, even if we ignore the above problems, all that the Rothbard-Mises analysis implies is that there is a tendency towards equilibrium *in a world with frozen data*. Of course, this implies little or nothing about whether there is a tendency towards equilibrium in a world where the data are not frozen. All that *ERE* theorists are saying is that, if we freeze the disequilibrating forces, then the equilibrating forces will prevail. But on this basis we may likewise assert a tendency towards *disequilibrium*. By allowing the data to change just as it does in the real world, and "freezing" all individual learning, we can demonstrate that the economy would degenerate into a series of successively less-coordinated states of disequilibrium. However, this would clearly be an illegitimate proof of a real world tendency towards disequilibrium, just as the Rothbard-Mises analysis does not succeed in establishing a real world tendency towards equilibrium.

The second function of the *ERE*—a starting point for analyzing the determination of market price (or other phenomena)—is also open to question. While it is desirable to analyze price determination in a simpler setting before examining the real world, it is doubtful whether the *ERE* is the proper initial setting, because *there are no prices in the ERE*. Prices are institutions that have evolved over time in order to help coordinate the plans of market participants. In a world in which all plans are already coordinated and actors possess all relevant information, prices would not serve any function.[6] The only prices that would exist in the *ERE* are the *ex post* barter ratios resulting from realized transactions, such as "ten apples for one horse." However, the notion of price as an *ex ante* disequilibrium institution which communicates knowledge (see Friedrich Hayek, 1945) has no place in the *ERE*.

Rothbard and Mises simply assume that prices exist in the *ERE* and then explain their determination. Interestingly, the contradictions involved with this assumption are

[3]Also see Frank Hahn (1972).

[4]Roman Frydman (1982) and George Richardson (1960) argue that learning processes will never converge upon an equilibrium. Existing "Austrian" models demonstrate convergence only for single markets under highly restrictive assumptions. See Jack High (1980, pp. 148–76); S. C. Littlechild and G. Owen (1980).

[5]Because the *ERE* does not stipulate a convex consumption set, people may have tastes for new and unique events or technologies where the "newness" or "uniqueness" is valued per se, thus making the initial "tastes freeze" self-contradictory. Likewise, the notion of freezing resources or resource patterns makes sense only in a world with no disequilibrium production or consumption to disturb the ability of physical processes to reproduce themselves.

[6]See John Hicks (1967, p. 3) who argues that numeraire prices are not even used for accounting purposes by *GE* (and presumably, *ERE*) market participants.

not denied; Rothbard and Mises emphasize the contradictory nature of the *ERE*, but still stress its usefulness. However, by forcing prices into a framework (the *ERE*) where they cannot logically exist, we must inevitably distract attention from the disequilibrium features of prices that characterize the real world. Even the static allocative function of prices is misrepresented by the *ERE*, since the allocative efficiency of the system is achieved by actors' knowledge of the external data (the "real factors") rather than by their observations of prices. How can an imaginary construct illuminate an institution that performs absolutely no function within that construct?

The *ERE* is also used as a starting point for analyzing the effects of an exogenous shock upon the economic system. Much of the analysis of the preceding section is applicable to our critique of this use of the *ERE* as well. The most important question for the economist studying the effects of an exogenous shock is how market institutions (prices, interest rates, firms, etc.) will deal with this change and how, in turn, this change will affect these institutions. However, a model with no institutions (or "make believe" institutions that serve no real function) is unsuitable for this task.

Depending upon the nature of the problem, introducing a change into an equilibrium setting may either understate or overstate the coordination problems that a market economy faces. Since all other activities are already coordinated, the disruptive nature of the change may be understated since the actors trying to deal with this change need not worry about transactions costs or plan conflicts among themselves. Furthermore, since the bonds that link individual plans in a disequilibrium world (such as *ex ante* prices) do not exist in the *ERE*, then the disruptive effects of the change upon these bonds will be either ignored or misrepresented.

The other possibility has equally serious analytical consequences. In equilibrium, many of the institutions that help coordinate behavior in a world of uncertainty are absent. Thus, we are giving the market economy an unfair test. We are throwing a new change into a system that, because it has anticipated all of the old changes, has no means for coping with the new change. In addition, many of the exogenous shocks introduced into the *ERE* also contradict the initial equilibrium conditions. For instance, using the *ERE* as a starting point for an analysis of monetary intervention (for example, Rothbard and Mises' business cycle theory) involves the contradiction of superimposing an increase in the money supply upon an essentially moneyless world.

The fourth proposed use of the *ERE*, as a "foil," seems to command the broadest support among *ERE* theorists. However, using the *ERE* as a foil contradicts the first three suggested uses of the *ERE*. If it is possible to look at the *ERE* in order to see what a price *is not*, what an interest rate *is not*, then this should be the *only* possible use for the *ERE*. If, as Mises claims (p. 348), the *ERE* has no human action, then we cannot claim there is a tendency towards equilibrium, since this would imply the nonsensical conclusion that there is a tendency for human action (and human institutions) to disappear. Nor could it be desirable to analyze price determination in the *ERE* as an "initial step" since our *ERE* foil tells us that there is neither price nor action in equilibrium. Finally, if the *ERE* reflects everything that the real world is not, introducing a change into the *ERE* and letting it work its way through the system cannot be a promising endeavor. At best, all such a procedure could be used for is to tell us how the real world does *not* react to change. Yet, when proponents of the *ERE* use the construct to analyze change, it is no longer viewed as a foil; instead it is looked upon as the initial "state of rest" that change acts upon. This is a clear inconsistency—the *ERE* either can be an initial state for an economy that is about to experience change, or it can be a foil, but it cannot be both. The very arguments made by Mises for using the *ERE* as a foil explain why it is of little value for serving any other purposes:

> ...[I]n the evenly rotating economy there is no choosing and the future is not uncertain as it does not differ from the present known state. Such a rigid

> system is not peopled with living men making choices and liable to error; it is a world of soulless unthinking automatons; it is not a human society, it is an ant hill. [p. 248]

However, the *ERE* is a poor foil because it is neither completely unrealistic nor internally consistent.[7] If one desires to use equilibrium as a foil, Debreu's model is superior to the *ERE*. The Debreu model is completely and consistently unrealistic: it has no money, no sequential transactions, perfect futures markets, and all of the other extreme assumptions necessary to accurately illustrate the other worldliness of the equilibrium concept. Nevertheless, regardless of one's opinion of the usefulness of the Debreu foil, the *ERE* is neither a useful substitute for, nor complement to, the standard *GE* model.

[7] The most important inconsistencies in the *ERE* are the existence of money and money prices and the absence of perfect futures markets. The first inconsistency tends to generate "neutrality of money" results, while the second leads to an underestimation of the difficulty of reasonably tight interpersonal coordination.

REFERENCES

Debreu, Gerald, *Theory of Value*, New Haven: Yale University Press, 1959.

Fisher, Franklin, "The Stability of General Equilibrium: Results and Problems," in M. Artis and A. Nobay, eds., *Essays in Economic Analysis*, New York: Cambridge University Press, 1976.

Frydman, Roman, "Toward an Understanding of Market Processes: Individual Expectations, Learning, and Convergence to Rational Expectations Equilibrium," *American Economic Review*, September 1982, *72*, 652–82.

Garrison, Roger, "Austrian Macroeconomics," in Louis Spadaro, ed., *New Directions in Austrian Economics*, Kansas City: Sheed, Andrews, and McMeel, 1978.

Hahn, Frank, "On the Foundations of Monetary Theory," in M. Artis and A. Nobay, eds., *Essays in Modern Economics*, New York: Barnes and Noble, 1972.

______, *On the Notion of Equilibrium in Economics*, New York: Cambridge University Press, 1973.

Hayek, Friedrich, "The Use of Knowledge in Society," in *Individualism and Economic Order*, Chicago: University of Chicago Press, 1945.

Hicks, John, *Critical Essays in Monetary Theory*, New York: Oxford University Press, 1967.

High, Jack, "Maximizing, Action, and Market Adjustment," unpublished doctoral dissertation, UCLA, 1980.

Littlechild, S. C, and Owen, G., "An Austrian Model of the Entrepreneurial Market Process," *Journal of Economic Theory*, December 1980, *23*, 361–79.

Mises, Ludwig, *Human Action*, New Haven: Yale University Fress, 1949.

Richardson, George B., *Information and Investment*, New York: Oxford University Press, 1960.

Rothbard, Murray, *Man, Economy, and State*, Los Angeles: Nash Publishing, 1962.

White, Larry, "The Austrian School and Spontaneous Order: Comment on O'Driscoll," *Austrian Economics Newsletter*, Spring 1979.

______, "Mises, Hayek, Hahn, and the Market Process: Comment on Littlechild," in Israel Kirzner, ed., *Method, Process, and Austrian Economics*, New York: Lexington Books, 1932.

Price Discrimination and Social Welfare

By HAL R. VARIAN*

The effect on social welfare of third-degree price discrimination was first investigated by Joan Robinson (1933). Richard Schmalensee (1981) has recently reexamined this question and presented several new results. In particular, he noted that a *necessary* condition for price discrimination to increase social welfare—defined as consumers' plus producers' surplus— is that output increase.

Schmalensee established this result only in the case of independent demands and constant marginal costs. However, it turns out to be true in much more general circumstances. In this paper I show how simple methods from duality theory can be used to establish this result and several other new results on the welfare effect of price discrimination.

I. A Reservation Price Model

Before proceeding to an examination of price discrimination in a general context, it is worth pausing to consider the special case of a reservation price model. I will describe the model in the context of discrimination by age—as in senior citizen discounts or youth discounts—but several other interpretations are possible. Assume that we have a set of consumers of different ages, and that one unit will be demanded by the consumers of age a if the price facing these consumers, $p(a)$, is less than or equal to $r(a)$, the reservation price of these consumers. Suppose that the slope of $r(a)$ is of one sign, which without loss of generality we take to be negative. For simplicity, it is assumed that costs are zero, or equivalently, that constant marginal costs are incorporated into the definition of $r(a)$.

Suppose first that the monopolist must choose one price p_0 that will apply to all consumers. Then the maximization problem facing the monopolist is to chose a_0 to solve:

$$\max r(a_0)a_0.$$

Now suppose that the monopolist is allowed to price discriminate; that is, he can choose critical ages a_1, a_2 and prices p_1, p_2 such that the consumers younger than a_1 face price p_1 and consumers between a_1 and a_2 face price p_2. The problem facing the monopolist now is to solve:

$$\max r(a_1)a_1 + r(a_2)(a_2 - a_1).$$

In this model it is easy to see that consumers' plus producers' surplus is given by the area below the reservation price function, as depicted in Figure 1. Thus the total welfare rises when price discrimination is allowed if and only if total output goes up. And, as shown below, output must always rise in this sort of model.

FACT 1: *If $r(a)$ is a decreasing function, then output and thus welfare must increase when price discrimination is allowed.*

PROOF:

Assume not so that $a_0 > a_2$ and thus: $-r(a_0)a_1 > -r(a_2)a_1$. By profit maximization: $r(a_0)a_0 \geq r(a_2)a_2$. Adding these two inequalities together, and adding $r(a_1)a_1$ to each side of the resulting inequality gives

$$r(a_1)a_1 + r(a_0)(a_0 - a_1) > r(a_1)a_1 + r(a_2)(a_2 - a_1),$$

which contradicts profit maximization.

This result easily generalizes to the choice of many regimes of price discrimination as well: allowing more price discrimination al-

*Department of Economics, University of Michigan, Ann Arbor, MI 48109. This research was supported in part by the National Science Foundation. I thank Richard Schmalensee, Louis Phlips, Andreu Mas-Collel, and an anonymous referee for helpful comments on an earlier draft.

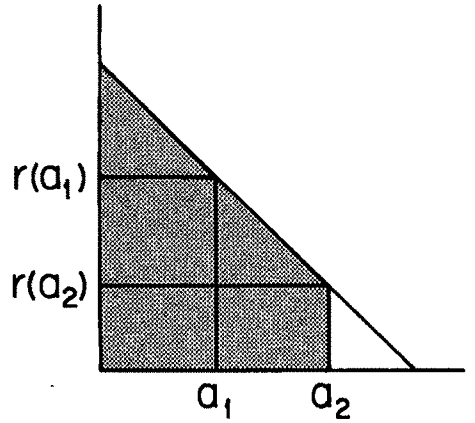


FIGURE 1. SURPLUS IN RESERVATION PRICE MODEL

ways increases output and welfare. As the number of prices increases to infinity, we converge to perfect price discrimination and thus maximal social welfare.

In this model we have a very simple story about price discrimination: price discrimination always increases output and an increase in output is always associated with an increase in welfare. But the reservation price model is a very special sort of demand structure and it is worth investigating whether these results carry over to more general demand specifications. As Schmalensee shows, in general, output and welfare may increase or decrease when price discrimination is allowed, although an increase in output remains a necessary condition for welfare increase. This result provides an observable criterion for when welfare has gone down under price discrimination, but how can we recognize those circumstances in which welfare has increased? I provide some answers to this question and related questions below.

II. Quasi-Linear Utility and Consumers' Surplus

I want to continue to use the classical measure of consumers' plus producers' surplus, and the most general preference structure for which that is possible is that of quasi-linear utility, which is also known as the case of "constant marginal utility of income." For this class of preferences it is well known that not only does consumer's surplus serve as a legitimate measure of individual welfare, but also that the individual consumers' utility functions can be added up to form a social utility function, so that aggregate *consumers'* surplus is also meaningful. For a discussion of consumers' surplus and indirect utility, see my 1984 book (ch. 7). These observations imply that we can treat the aggregate demand function as though it were generated by a representative consumer with an indirect utility function of the form:

$$V(\mathbf{p}, y) = v(\mathbf{p}) + y.$$

The aggregate consumer's income, y, is composed of some exogenous income which we take to be zero and the profits of the firm. Thus the appropriate form of the social objective function becomes:

$$V(\mathbf{p}, y) = v(\mathbf{p}) + \pi(\mathbf{p}).$$

By Roy's law the demand for good i is given by the negative of the derivative of $v(\mathbf{p})$ with respect to p_i—since the marginal utility of income is one. Thus the integral of demand is just $v(\mathbf{p})$. It follows that the above expression is nothing but the classical welfare measure of consumers' plus producers' surplus.

As a general principle, it is easier to differentiate to find demands than to integrate to find surplus; thus starting with the properties of the indirect utility function rather than the demand functions tends to simplify most problems in applied welfare economics. The most important property for our purposes concerns the curvature of the indirect utility function. The indirect utility function is always a quasiconvex function of prices, but in the case of quasi-linear utility, it is not hard to show that it is in fact a *convex* function of prices. (Proof: the expenditure function is $e(\mathbf{p}, u) = u - v(\mathbf{p})$ and it is necessarily a concave function of prices.)

III. Upper and Lower Bounds on Welfare Change

I turn now to the welfare effects of price discrimination for demand structures generated by quasi-linear utility. I start by describ-

ing a general result about such demands which can then be specialized in a number of ways. Consider an initial set of prices $\mathbf{p}^0$ and a final set of prices $\mathbf{p}^1$, and let $c(\mathbf{x}(\mathbf{p}^0))$ and $c(\mathbf{x}(\mathbf{p}^1))$ denote the total costs of production at the two different output levels associated with the price vectors $\mathbf{p}^0$ and $\mathbf{p}^1$. Let $\Delta\mathbf{x}$ denote the vector of changes in demand (i.e., $\Delta\mathbf{x} = \mathbf{x}(\mathbf{p}^1) - \mathbf{x}(\mathbf{p}^0)$), and let Δc denote the change in the total costs of production.

FACT 2: *The change in welfare, ΔW, satisfies the following bounds*:

$$\mathbf{p}^0\Delta\mathbf{x} - \Delta c \geq \Delta W \geq \mathbf{p}^1\Delta\mathbf{x} - \Delta c.$$

PROOF:

Since the indirect utility function is a convex function of prices, we have:

$$v(\mathbf{p}^0) \geq v(\mathbf{p}^1) + \mathbf{D}v(\mathbf{p}^1)(\mathbf{p}^0 - \mathbf{p}^1)$$

where $\mathbf{D}v(\mathbf{p})$ stands for the gradient of $v(\mathbf{p})$. Using Roy's law, and rearranging:

$$\mathbf{x}(\mathbf{p}^1)(\mathbf{p}^0 - \mathbf{p}^1) \geq v(\mathbf{p}^1) - v(\mathbf{p}^0) = \Delta v.$$

The change in profits is given by

$$\mathbf{x}(\mathbf{p}^1)\mathbf{p}^1 - \mathbf{x}(\mathbf{p}^0)\mathbf{p}^0 - \Delta c = \Delta\pi.$$

Adding these expressions together we have

$$[\mathbf{x}(\mathbf{p}^1) - \mathbf{x}(\mathbf{p}^0)]\mathbf{p}^0 - \Delta c$$
$$= \mathbf{p}^0\Delta\mathbf{x} - \Delta c \geq \Delta v + \Delta\pi = \Delta W.$$

The other bound can be derived in a similar manner.

Now think of the n goods as being one good sold in n different markets and produced at constant marginal cost. I want to compare a uniform pricing policy to a policy of price discrimination. Making the necessary substitutions in the bounds given in Fact 2, we have the following:

FACT 3: *Let* $\mathbf{p}^0 = (p_0, \ldots, p_0)$, $\mathbf{p}^1 = (p_1, \ldots, p_n)$, *and let c be the constant level of marginal costs. Then the bounds on welfare change become*

$$(p_0 - c)\sum_{i=1}^{n} \Delta x_i \geq \Delta W \geq \sum_{i=1}^{n} (p_i - c)\Delta x_i.$$

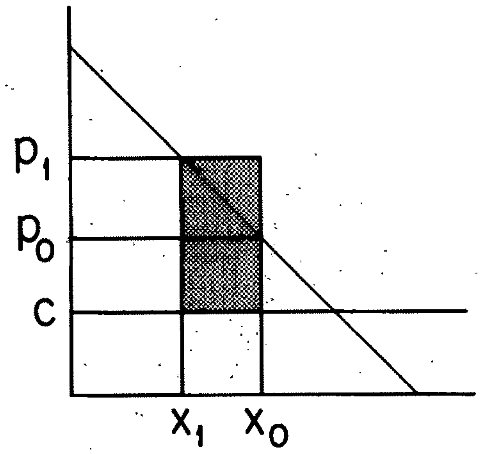


FIGURE 2. BOUNDS ON WELFARE CHANGE IN SINGLE MARKET

Note that the upper bound in Fact 3 immediately gives Schmalensee's result that an increase in output is a necessary condition for welfare to increase. The lower bound in Fact 3 was not discussed by Schmalensee. It implies that if the profitability of the new output exceeds the profitability of the old output, *valued at the new prices*, then welfare must have risen at the discriminatory equilibrium. This is basically a revealed preference relationship.

Both of these facts hold in complete generality, for independent and dependent demands, as long as one is willing to assume quasilinear utility; that is, that aggregate consumers' surplus serves as an acceptable welfare measure. The bounds have a simple geometric interpretation in the case of a single demand curve which is given in Figure 2. However, it is worth emphasizing that these results are purely statements about demand and utility functions and hold for arbitrary configurations of prices. The fact that the prices are chosen by a profit-maximizing monopolist has not been used in their derivation.

IV. Bounds on Welfare Change with Optimal Price Discrimination

I now ask what results can be derived that *use* the conditions implied by *profit-maximizing* price discrimination. Let us specialize the notation above to consider only three prices, the initial price p_0 that is charged in both markets, and the final prices p_1 and p_2 that are profit-maximizing prices in their respective markets. We also continue to suppose that the good is produced at constant marginal cost c.

Fact 3 holds for all prices and all demand structures. If we consider only profit-maximizing prices and restrict ourselves to the textbook case of independent demands, we can apply the standard marginal revenue equals marginal cost formulas to find:

FACT 4: *If demand functions are independent, welfare is bounded by*

$$\frac{c[\Delta x_1+\Delta x_2]}{\varepsilon_0-1}\geq \Delta W\geq \frac{c\,\Delta x_1}{\varepsilon_1-1}+\frac{c\,\Delta x_2}{\varepsilon_2-1},$$

where $\varepsilon_0, \varepsilon_1, \varepsilon_2$ are the (absolute values of the) respective elasticities of demand, evaluated at p_0, p_1, and p_2.

This result may be of use if one has estimates of the elasticities of demand in the various submarkets. However the independent demand case is rather restrictive. Profit maximization *alone* yields the following sufficient condition for a welfare increase.

FACT 5: *A sufficient condition for welfare to increase under profit-maximizing price discrimination is that*

$$(p_0-c)[x_1(p_0,p_0)+x_2(p_0,p_0)]$$
$$>(p_1-c)x_1(p_0,p_0)+(p_2-c)x_2(p_0,p_0).$$

PROOF:

By profit maximization at (p_1, p_2) we have

$$(p_1-c)x_1(p_1,p_2)+(p_2-c)x_2(p_1,p_2)$$
$$\geq(p_0-c)x_1(p_0,p_0)+(p_0-c)x_2(p_0,p_0).$$

Combining this with the hypothesis and rearranging, we have $(p_1-c)\Delta x_1+(p_2-c)\Delta x_2>0$. By Fact 3 this yields a welfare increase.

The interesting thing about Fact 5 is that it only involves a condition on the nondiscriminatory levels of output. If you can forecast the prices that would be charged under discrimination and those prices satisfy the condition given in Fact 5, you can be assured that welfare will rise when discrimination is allowed.

It might be worthwhile to give an example of how these bounds can be used to verify that a welfare increase or decrease has occurred. The simplest example is the case of linear demands described by Schmalensee. If both markets are served in the single price regime, then it is easy to show by direct calculation that total output with discrimination is the same as in the single price regime. Hence, as noted by Schmalensee, welfare must decline when discrimination is allowed.

However, suppose we are in a situation where market 2 is not served in the single price regime. Then when discrimination is allowed, $p_1=p_0$, $\Delta x_1=0$, and $\Delta x_2>0$. By Fact 3 welfare must increase. Note also that in this situation the sufficient condition given in Fact 5 is satisfied as an equality.[1]

Thus Fact 3 verifies that welfare will increase when price discrimination is allowed in the linear demand case if a new market is served. However, Fact 3 also shows that for *arbitrary* independent demands, welfare goes up if a new market is served when price discrimination is allowed. The argument is simply that of the above paragraph: $\Delta x_1=0$ and $\Delta x_2>0$, so welfare must increase.

These examples give some intuition for the case where both markets are served in both the discriminatory and nondiscriminatory regimes as in Figure 4. What is needed for welfare to increase when price discrimination

[1] Of course, total output rises as well. The reader might wonder what is wrong with the "direct calculation" mentioned above. The problem is that what economists call "linear" demand curves are not really linear functions; instead they have the form: $Q=\max\{A-BP,0\}$.

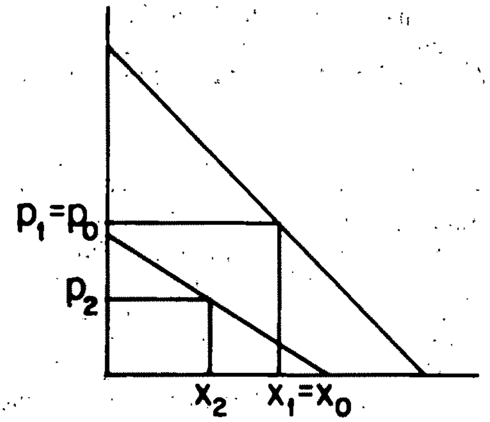


FIGURE 3. INCREASE IN WELFARE (BOUNDARY CASE)

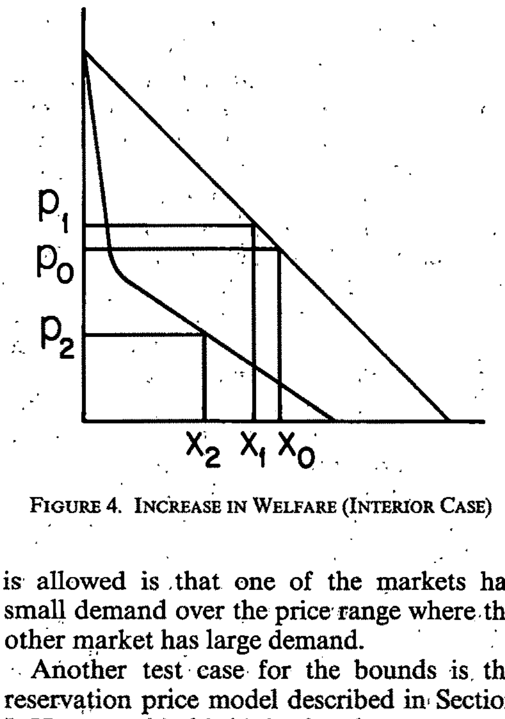


FIGURE 4. INCREASE IN WELFARE (INTERIOR CASE)

is allowed is that one of the markets has small demand over the price range where the other market has large demand.

Another test case for the bounds is the reservation price model described in Section I. Here we should think of each consumer as being a different market with demand function $x_a(p)$. If there are a_0 consumers purchasing the good in the single-price regime and $a_2 > a_0$ under price discrimination, then we know that $\Delta x_a = 0$ for $a \leq a_0$ and $\Delta x_a = 1$ for $a_2 \geq a > a_0$, which by Fact 3 implies welfare must increase when discrimination is allowed.

The bounds can also be used to show that marginal cost pricing and perfect price discrimination are welfare optima in the reservation price model. For if price equals marginal cost, the upper bound on welfare change is zero. And if each consumer is being charged his reservation price, then Δx_a is either 0 or -1 which implies the upper bound in nonpositive.

The welfare bounds given above take a nice form if we are willing to make curvature assumptions on the demand functions. Let us restrict ourselves to the case of independent demands and focus on the market for good 1. Then the argument of Fact 2 implies that the welfare effect of a price change of good 1 is bounded by $(p_0 - c)\Delta x_1 \geq \Delta W_1 \geq (p_1 - c)\Delta x_1$. Suppose that the demand for good 1 is a concave function of its own price. Then we have $\Delta x_1 \geq x_1'(p_1)(p_1 - p_0)$. Combining these two inequalities we have $\Delta W_1 \geq (p_1 - c)x_1'(p_1)[p_1 - p_0]$. The first-order conditions for profit maximization imply that $(p_1 - c)x_1'(p_1) + x_1(p_1) = 0$. Substituting we have $\Delta W_1 \geq x_1(p_1)(p_0 - p_1)$. If both markets have concave demand curves we can write:

$$\begin{aligned}\Delta W &\geq x_1(p_1)(p_0 - p_1) + x_2(p_2)(p_0 - p_2) \\ &= p_0[x_1(p_1) + x_2(p_2)] \\ &\quad - [p_1 x_1(p_1) + p_2 x_2(p_2)].\end{aligned}$$

Add and subtract $(p_0 - c)[x_1(p_0) + x_2(p_0)] - c[x_1(p_1) + x_2(p_2)]$ to get $\Delta W \geq (p_0 - c)\Delta x - \Delta\pi$, where Δx is the total change in output and $\Delta\pi$ is the total change in profits. Thus the change in welfare is at least as large as the change in profit valued at the old prices minus the change in actual profit. Or, to put it another way, $\Delta x > \Delta\pi/(p_0 - c)$ is a sufficient condition for welfare to increase when price discrimination is allowed if all demand curves are independent and concave. Combining this with Fact 3 we can conclude:

FACT 6: *If all demand curves are independent and concave the welfare bounds can be written as*

$$(p_0 - c)\Delta x \geq \Delta W \geq (p_0 - c)\Delta x - \Delta\pi.$$

Note that Facts 5 and 6 use profit maximization at p_1 and p_2, but do *not* use profit maximization at p_0. Thus these results are independent of firm behavior at the nondiscriminatory equilibrium.

If the demand curves are concave and convex (i.e., linear), then the inequality in Fact 6 becomes an equality so that $\Delta W = -\Delta\pi$. Thus in the case of linear demands, the change in welfare is exactly the negative of the change in profits. Of course this can also be verified by direct calculation.

V. More General Cost Structures

The above results were all derived in the case of constant marginal cost but they can be partially extended to the case of increasing marginal costs; that is, the case of a *convex* cost function. By the standard convexity inequality:

$$\mathbf{Dc}(\mathbf{x}(\mathbf{p}^1))\,\Delta\mathbf{x} \geq \Delta c \geq \mathbf{Dc}(\mathbf{x}(\mathbf{p}^0))\,\Delta\mathbf{x}.$$

Combining this with the inequality given in Fact 2 we have

$$[\mathbf{p}^0 - \mathbf{Dc}(\mathbf{x}(\mathbf{p}^0))]\,\Delta\mathbf{x} \geq [\mathbf{p}^1 - \mathbf{Dc}(\mathbf{x}(\mathbf{p}^1))]\,\Delta\mathbf{x}.$$

Again, these are general bounds which hold for all pairs of price vectors $\mathbf{p}^0$ and $\mathbf{p}^1$ as well as for arbitrary convex cost functions; in particular the cost function can be a function of the vector of outputs rather than just the total output. Thus the bounds can be useful in more general contexts. For example, they give a simple proof of the optimality of marginal cost pricing in the presence of convex costs: if $\mathbf{p}^0 = \mathbf{Dc}(\mathbf{x}(\mathbf{p}^0))$ then any movement from $\mathbf{p}^0$ must decrease social welfare.

If costs depend only on total output, denoted by x_0 and x_1, and $\mathbf{p}^0$ is a vector of constant prices p_0 as above, we can write these bounds as

$$[p_0 - c'(x_0)] \sum_{i=1}^{n} \Delta x_i \geq \Delta W \geq \sum_{i=1}^{n} [p_i - c'(x_1)]\,\Delta x_i.$$

Thus in the case of increasing marginal costs, Schmalensee's proposition still holds: price must be greater than marginal cost at the nondiscriminatory price, so an increase in output is still a necessary condition for welfare to increase.

REFERENCES

Robinson, Joan, *Economics of Imperfect Competition*, London: Macmillan, 1933.

Schmalensee, Richard, "Output and Welfare Implications of Monopolistic Third-Degree Price Discrimination," *American Economic Review*, March 1981, *71*, 242–47.

Varian, Hal R., *Microeconomic Analysis*, 2d ed., New York: W. W. Norton, 1984.

Controlling Contradictions Among Regulations: Note

By CHRIS T. HENDRICKSON AND FRANCIS CLAY MCMICHAEL*

Single-minded pursuit of simple objectives can be effective within the narrow terms of the objective, but it may also create some unwanted effects. In his 1984 article, Lester Lave noted that Congress typically pursues externalities with a "one at a time" strategy. He suggested that both the legislature and regulatory agencies would be better served by considering the general cost effects of specific regulations. For example, regulations requiring air pollutant emissions controls on motor vehicles may decrease vehicle safety (via induced changes in vehicle weights), or increase fuel consumption (with undesirable external effects due to increased aggregate demand for petroleum), as well as increase costs to producers or car owners. Lave proposed that the benefits of reduced costs associated with correcting single externalities be weighted against the resulting additional costs in formulating regulations.

We suggest an extension to Lave's argument and analysis. In essence, we note that physical or technical constraints often exist on the production and control of particular unwanted by-products. In effect, irreducible pollutants must be discharged to one or more of the air, water, or land media. Each disposal media involves costs of various sorts. Maximization of social welfare becomes a constrained problem, and regulatory policies should consider cross-media effects in such cases.

Economic activities are similar in one essential way. They take in streams of materials and energy and produce desired products and/or useful energy. They also produce material effluents and waste energy. The conservation laws of physics require that material and energy inputs and outputs balance (R. U. Ayres, 1977). Large-scale material balance models are becoming increasingly important as we attempt to locate all of the places where polluting substances appear in industrial processes, and develop strategies for substitutions and for material tracking, management, and control (M. G. Morgan and McMichael, 1981). Environmental regulatory agencies, both at federal and local levels, are typically concerned with specific media (for example, air, land or water) and legislation is often specific to a single media.

An often cited goal is zero discharge of pollutants into one media. In practice, this goal typically results in the transfer of pollutants from one media to another. For example, the Commonwealth of Pennsylvania prohibited discharge of phenolic-bearing coke plant wastewater into rivers in 1929 after the development of a coke-quenching practice which volatized the pollutants to the air. This new practice switched the pollutant discharge from one media to another. Due to increased concern over air pollution, the use of wastewater for coke quenching was prohibited in Allegheny County, Pennsylvania, in 1970 unless the wastewater "was of a quality as may be discharged to the nearest receiving stream or river, in accordance with the Acts of the Commonwealth of Pennsylvania." In effect, both water and air emissions were forbidden. The two conflicting environmental regulations were only resolved when entirely new treatment technologies were developed for coke plant wastewaters (R. W. Dunlap and McMichael, 1971; 1976). Typically, minimum environmental impact on any given media can be established by diverting emissions to other media. The overall environmental impact is rarely considered and, indeed, its consideration is fraught with difficulties. Consideration of net environmental impact begins with the realization that single-media solutions are rarely optimal.

Many cross-media effects might be captured by a simple constraint which requires

*Associate Professor of Civil Engineering, and Professor of Civil Engineering and Engineering Public Policy, respectively, Carnegie-Mellon University, Pittsburgh, PA 15213.

that emissions to the air, water, or land must equal the sum of the pollutants from a production process. This linear constraint could represent a mass or energy balance for a single species such as heat, total dissolved solids, or chlorides in wastewater discharges. The constraint simply implies that, say, wastewater chlorides *must* go somewhere; they do not disappear. With a mass balance constraint of this sort, Lave's model of welfare maximization would become a problem of constrained optimization. The first-order optimality conditions would then include an additional term representing the Lagrangean multiplier on the balance constraint. This multiplier can be interpreted as a "shadow price" of production or the social cost of additional pollutant emissions. Also, the marginal social cost of emission control for each media would be set equal to this shadow price at the optimum, implying that some emissions to each media would occur with well-behaved cost functions. In contrast, Lave's unconstrained model does not include the shadow price term, and emissions into each media would be chosen so that the total marginal social cost was zero.

This simple case of a mass balance constraint on a single species might be disaggregated into different locations or by other characteristics (for example, water emissions by stream location). More complicated mass balance constraints can also be formulated for chemical reactions and other technical processes. For example, emissions control at a local site may require emissions elsewhere to generate the energy required for the control process.

Both Lave's and our analysis suggest that a broader consideration of social costs be adopted in formulating regulatory policies. We suggest that material and energy balance constraints make such cross-media considerations particularly important. To recognize that emissions controls to one media will result in greater emissions elsewhere is an important aspect of such cost-estimation and evaluation procedures. Particularly in the area of environmental regulation, these constraints are common.

REFERENCES

Ayres, R. U., *Resources, Environment and Economics: Applications of the Materials/Energy Balance Principle*, New York: Wiley-Interscience, 1977.

Lave, Lester B., "Controlling Contradictions Among Regulations," *American Economic Review*, June 1984, *74*, 471–75.

Morgan, M. G. and McMichael, F. C., "A Characterization and Critical Discussion of Models and Their Use in Environmental Policy," *Policy Sciences*, June 1981, *13*, 345–70.

Dunlap, R. W. and McMichael, F. C., "Cross-Media Effects of Environmental Regulation: Technology Selection Considering Social Values," in *Proceedings*, Fifth National Conference of the Air Pollution Control Division, American Society of Mechanical Engineers, Pittsburgh, May 1971.

______ **and** ______, "Reducing Coke Plant Effluent," *Environmental Science and Technology*, No. 7, 1976, *10*, 654–57.

Fair Outcome/Fair Process

By ROBERT PIRON*

In a recent article, Randall Holcombe (1983) takes William Baumol (1982) to task for defining "fairness" as outcome independent of process, claiming, "Fair outcomes are the result of fair processes, and no theory can proclaim an outcome to be fair without examining the process that produced the outcome" (p. 1153). To illustrate his point he provides what he believes to be an obviously "fair" process: "Most people would agree that if one person builds two chairs in a day when an equally able person spends the day watching television, it would be fair for the first person to keep the two chairs..." (p. 1153).

This position is remarkably similar to that taken by Robert Nozick:

> A distribution is just if it arises from another (just) distribution by legitimate means. ... As correct rules of inference are truth preserving and any conclusion deduced via repeated application of such rules from only true premises is itself true, so the means of transition from one situation to another specified by the principle of justice in transfer are justice preserving, and any situation actually arising from repeated transitions in accordance with the principle from a just situation is itself just. [1973, p. 48]

The main point of this comment is to show that Holcombe has steered economists (inadvertently?) into a methodological morass they have tried hard and long to avoid for half a century, and to begin to frame the issue, consider Richard Bellman's Principle of Optimality: "An optimal policy has the property that whatever the initial state and initial decision, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision" (1957, p. 83). So, if Holcombe were dealing with "optimal" decisions in the Bellman context (dynamic Markov algorithms), he could correctly infer that "optimal outcomes are the result of optimal processes." Why, then, are we uneasy with "fair outcomes are the result of fair processes?" The answer seems obvious: optimal is, in this context, a mathematical (i.e., logical) construct, while fair is a value construct. The latter sets the moral juices flowing while the former does not. Since fair has been one of the slipperiest terms in ethics since the dawn of Moral Philosophy, economists should minimize *hubris* and let philosophers (and public choice theorists) battle things out on the appropriate turf and mind their own methodological business.

Here is a minitour of the economics literature on this subject:

> Now in so far as the idea of rational action involves the idea of *ethically appropriate* action, and it certainly is sometimes used in this sense in everyday discussion, it may be said at once... that no such assumption enters into economic analysis.... Economic analysis is *wertfrei* in the Weber sense. The values of which it takes account are valuations of individuals. The question whether in any further sense they are *valuable* valuations is not one which enters into its scope. If the word rationality is to be construed as in any way implying this meaning, then it may be said that the concept for which it stands does not enter into economic analysis. [Lionel Robbins, 1952, p. 91]

> Positive economics is in principle independent of any particular ethical position or normative judgments.... Its performance is to be judged by the precision, scope, and conformity with experience of the predictions it yields. In short, positive economics is, or can

*Department of Economics, Oberlin College, Oberlin, OH 44074. Thanks go to Lloyd Orr, Albert MacKay, David Cleeton, Luis Fernandez, and an anonymous referee for helpful comments and references.

> be, an "objective" science, in precisely the same sense as any of the physical sciences.... If this judgment is valid, it means that a consensus on "correct" economic policy depends much less on the progress of normative economics proper than on the progress of a positive economics yielding conclusions that are, and deserve to be, widely accepted. It means also that a major reason for distinguishing positive economics sharply from normative economics is precisely the contribution that can thereby be made to agreement about policy.
> [Milton Friedman, 1953, pp. 4–7]

> In the past few years, several economists... have produced a novel analytical theory of *fairness* in the distribution of resources, in contradistinction to the efficiency of their allocation. This work is primarily *philosophical* [emphasis added] in orientation, being concerned primarily with the logical underpinnings of an analysis of fair division, rather than with its application. [Baumol, p. 639]

> It was to be expected, I suppose, that Simon Blackburn's elegant review ...should end in a plea for continued employment of Philosophers. But we have a useful piece of jargon in economics that argues against it, "negative externalities"—like smoke from the local mill. The activities of Philosophers are no trouble at home perhaps, but spill over into neighboring places. The claim of Philosophy to be a meta-science is a public nuisance.... The neighbors have suffered quite enough from the impulse to Philosophize about good reasons in politics, economics, or law.
> [Donald McCloskey, 1983, p. 882]

Now, ignoring, seemingly, both this literature and the teflon characteristics of fair, Holcombe confidently criticizes Baumol by arguing that the economics definition of fair must include the fairness of the process underlying any analysis of the problem, plunging us directly into a type of intellectual exercise that, like the cosmos in some views, has neither beginning nor end. To illustrate: what Holcombe holds as rather obvious, namely that people are entitled to the fruits of their labor, and any process that allows that is fair, is, in fact, remarkably unobvious and really terribly controversial. For example, in a review of Rawls, Kenneth Arrow has said,

> The general point of view [of *A Theory of Justice*] is a strong affirmed egalitarianism, to be departed from only when it is in the interest of all to do so. "All social values—liberty and opportunity, income and wealth, and the bases of self-respect—are to be distributed equally unless unequal distribution of any, or all, of these values is to everyone's advantage." This *generalized difference principle*, as Rawls terms it, is no tautology. In particular it implies that even natural advantages, superiorities of intelligence or strength, do not in themselves create any claim to greater rewards. The principles of justice are an agreement to regard the distribution of natural talents as a common asset and to share in the benefits of this distribution.
>
> Personally, I share fully this value judgment; and, indeed, it is implied by almost all attempts at full formalization of welfare economics. But a contradictory proposition: that an individual is entitled to what he creates, is widely and unreflectively held; when teaching elementary economics, I have had considerable difficulty in persuading the students that this *productivity principle* was not completely self-evident. [1973, p. 247]

So Holcombe does the methodology of social science no service with his critique of Baumol, since his views on fair processes do not, and very likely in principle cannot, command the wide acceptance required by positive economics, being by their nature fundamentally normative and therefore infinitely arguable. Baumol, in his writing on fairness has tried valiantly to derive empirically meaningful propositions from his relatively straightforward and ethically neutral *analysis* of fairness, even though it can (and should) be argued that, as Geoffrey Philpotts (1983)

has shown, even Baumol's *definition* of fairness as "non-envy-producing division" is also ethically loaded, and who is to say it is not "as loaded" as Holcombe's? However, the point of this note is to show that it is methodologically futile in arguments concerning an economic model to supplant one ethical presupposition with another. Rather, one should demand, in this case, Holcombe of Baumol, that his (Baumol's) model pass predictive-positive tests, not ethical-normative ones.

REFERENCES

Arrow, Kenneth J., "Rawls' *A Theory of Justice*," *Journal of Philosophy*, May 10, 1973, *70*, 245–63.

Baumol, William J., "Applied Fairness Theory and Rationing Policy," *American Economic Review*, September 1982, *72*, 639–51.

Bellman, Richard E., *Dynamic Programming*, Princeton: Princeton University Press, 1957.

Friedman, Milton, *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.

Holcombe, Randall G., "Applied Fairness Theory: Comment," *American Economic Review*, December 1983, *73*, 1153–56.

McCloskey, Donald N., "Consequences of Pragmatism," *Times Literary Supplement*, August 19, 1983, p. 882.

Nozick, Robert, "Distributive Justice," *Philosophy and Public Affairs*, Fall 1973, *3*, 45–126.

Philpotts, Geoffrey, "Applied Fairness Theory: Comment," *American Economic Review*, December 1983, *73*, 1157–60.

Robbins, Lionel, *An Essay on the Nature and Significance of Economic Science*, 2d ed., London: Macmillan, 1952.

Another Look at Free Banking in the United States

By JAMES A. KAHN*

In their 1983 paper, Arthur Rolnick and Warren Weber document the experience of several states' experiments with free banking in the decades leading up to the Civil War, using data on bank openings and closings, and on redemption rates of failed banks (the rate, per dollar par value, that failed banks were able to pay their note holders). In this note I extend the Rolnick-Weber analysis by comparing the experiences of free and non-free-banking states with regard to the frequency of bank closings and failures, to give some perspective to their conclusion that free banks were not as short-lived as has been the conventional view. In so doing I offer an alternative set of statistics to those provided by Rolnick-Weber on bank life expectancies and average redemption rates of failed banks, statistics that I argue take better account of the truncation problems inherent in the data. I find that bank closings were much more frequent in free bank states. This was especially so in the first five years after deregulation, but remained true to some extent throughout the period.

I. Bank Failure Rates

One problem often attributed to the Free Banking Era, and to unregulated banking more generally, is a high rate of bank failures. As correctly pointed out by Rolnick-Weber, what is relevant from a public policy standpoint is not so much failure in the sense of going out of business, but rather more in terms of the inability to redeem notes at their par value. That is, the public policy concern is primarily the provision of a secure currency, not the safety of stockholders' investments. Nonetheless, because data on redemption rates of closed banks are difficult to find for many states, and because the closure rate should give some indication of the quality and stability of a state's banking system, I present here state-by-state comparisons of bank "life expectancies." The data come from the same sources as those used by Rolnick-Weber (*House Executive Documents*, 1837–63).

One difficulty in the estimation of life expectancies over a finite sample period is that many of the banks will still be in operation at the end of the period. Rolnick and Weber assign to such banks their age at the end of the sample period and then calculate the sample means. They recognize the obvious bias in such a procedure, and present their estimates as lower bounds. Their goal was to show that most banks were in business for more than just a few months or a year. Unfortunately, one cannot use such calculations to make meaningful state-to-state comparisons, because both the length of the sample period and the proportion of failed banks to the total number of banks varied tremendously across states.

I show below that under the assumption that the underlying stochastic process can be represented by a constant hazard rate for bank failures, and that the observations are independent, correction of the truncation bias is a simple matter. By assumption one can ignore the truncation at the start of the period, and pretend that all banks in operation at that time opened within the previous year. It then turns out that a consistent estimate of the life expectancy is just the total number of bank-years in the sample, divided by the number of failed banks, or, to put it another way, the probability of failure in a year is estimated by the total number of failures divided by the total number of bank-years. Besides simplicity, this procedure has other desirable features as well. One can get reasonable estimates even for relatively short sample periods (as will be seen below), and one can get consistent estimates of standard errors as well.

*Graduate student, Massachusetts Institute of Technology, Cambridge, MA 02139. I thank Peter Temin for comments, and A. J. Rolnick for supplying some of the data.

Again, the point of this correction is not so much to get better statistics for the free bank states (after all, if most of a state's banks did fail, my statistics will not differ very much from Rolnick-Weber's). Rather it is to get statistics that are comparable across states, even when the lengths of the sample periods vary widely. Of course, this is not without its problems too. The phenomenon of bank panics raises some doubts about the independence of observations, and one might expect the hazard rate to decline over time. I take account, to some extent, of the latter possibility in one of the sets of calculations below, but mainly it is hoped that this simple correction at least enables one to make state-by-state comparisons; that is, that the biases are not too different from one state to the next. This could not be the case with Rolnick-Weber's statistics for the reasons stated above.

Suppose that a bank currently in operation has every year a probability p of going out of business, and let X_i denote the total number of years that bank i operated. Then one may show that

$$(1) \qquad Pr(X_i = n) = p(1-p)^{n-1},$$

$$(2) \qquad Pr(X_i > n) = (1-p)^n,$$

$$(3) \qquad E(X_i) = 1/p.$$

Letting k denote the life expectancy of the bank, it is clear from (3) that an equivalent model may be expressed in terms of k, with $p = 1/k$.

Now suppose we observe a sample of N banks over time, but that we do not necessarily observe X_i. Instead we observe Y_i, where $Y_i = X_i$ if $X_i < T_i$, and $= T_i$ otherwise, where T_i is the number of years between the opening of bank i and the last year of the sample. For concreteness, suppose the first M banks in the sample are observed to go out of business, while all the others are still operating at the end of the sample period.

From equations (1) and (2), and assuming that the observations are independent, the likelihood function for the N observations of Y_i is then

$$M\log(p)+\log(1-p) \times \left\{ \sum_{i=1}^{M}(X_i-1)+\sum_{j=M+1}^{N}T_j \right\}.$$

This is maximized at $p^* = \{M^{-1}\Sigma_{i=1}^{N} Y_i\}^{-1}$, so that $k^* = M^{-1}\Sigma_{i=1}^{N} Y_i$. Also, it turns out that a consistent estimate of the variance of the estimator k^* is $M^{-1}k^*(k^*-1)$.

Table 1 gives results for the three states used by Rolnick-Weber (New York, Wisconsin, and Indiana), along with results for Minnesota (a free bank state not included in their mean life calculations because the sample period was too short), and for three other states: Maine and Maryland, which had no free banking laws, and New Jersey, which passed a free banking law midway through the sample period.[1] The results show strikingly that the banks in free bank states had much shorter life expectancies (i.e., much higher failure rates). The table also provides estimates of the standard errors, which indicate that the means are estimated fairly precisely.

One might argue that the foregoing analysis overlooks the fact that any major change in the regulation of an industry is bound to cause some upheaval, at least initially, so that it is misleading to characterize the higher failure rates as a permanent effect of the free bank laws. (The airline industry today is a case in point.) For this reason, Table 2 provides estimates for the free bank states beginning roughly five years after the initiation of the policy (excluding Minnesota, which had free banking for only five years in all, and New York, for which the estimates in

[1] I attempted to choose non-free-bank states that were as unexceptional and as broadly representative as possible. Many of the lightly populated western states had only one state bank for much of the period, and so would not have provided much information. On the other hand, states such as Massachusetts, with large, well-established banking systems, would probably have been unrepresentative of state-chartered banking systems in general (just as, I will claim later, New York was not a typical free bank state).

TABLE 1—BANK LIFE EXPECTANCIES

	ME	MD	NJ	Total		
A. Non-Free-Bank States						
Period	1837–63	1837–63	1837–51			
Banks at Start	55	21	25	101		
Openings	59	22	9	90		
Closings	46	11	10	66		
Mean Life	31.2	61.1	38.7	37.8		
	(4.5)	(18.3)	(12.1)	(4.6)		
	NJ	NY	WI	IN	MN	Total
B. Free Bank States						
Period	1852–63	1843–62	1853–62	1854–63	1859–63	
Banks at Start	–	49	8	30	2	89
Openings	51	400	132	73	14	670
Closings	25	160	80	86	11	362
Missing	11	48	4	30	1	94
Mean Life	9.4	22.1	7.3	3.2	3.1	12.9
	(1.8)	(1.7)	(0.8)	(0.3)	(0.8)	(0.7)

Note: Standard errors are shown in parentheses.

TABLE 2—BANK LIFE EXPECTANCIES AFTER INITIAL YEARS

	ME	MD	
A. Non-Free-Bank States			
Period	1854–63	1854–63	
Banks at Start	60	24	
Openings	28	13	
Closings	19	5	
Mean Life	37.1	60.4	
	(8.4)	(26.8)	
	NJ	WI	IN
B. Free Bank States			
Period	1856–63	1857–62	1858–63
Banks at Start	35	45	19
Openings	21	85	2
Closings	8	70	5
Mean Life	24.7	7.4	20.7
	(8.5)	(0.8)	(9.0)

Note: Standard errors are shown in parentheses.

Table 1 suffice). It also presents mean life estimates for Maine and Maryland based on comparable sample periods, to take account of the possibility that there was something different about the 1850's that might have led to higher failure rates. The results indicate that while there was overall some improvement in the free bank states after the initial five years, bank failure rates remained considerably higher than in the non-free-bank states.[2] In the latter, failure rates were remarkably stable throughout the era.

II. The Costs of Holding Bank Notes

As stated above, the goal of the states' bank regulatory policies was presumably a sound currency, and the consequent public confidence in the banking system, rather than the safety of bank shareholder's investments. A high rate of bank failures is a symptom of difficulties, but may not itself have been considered a serious cost from a public policy viewpoint.

Unfortunately, data on redemption rates of failed banks are not as easy to come by. The four free bank states documented by

[2] The improvements may have been due to the weeding out of incompetents, to improved enforcement procedures by the states, or to modifications of the free banking laws—it is not clear from the documents.

Rolnick-Weber appear to be the only ones that included redemption rates on notes of failed banks in a systematic way in their reports to the Treasury Department. Therefore in this section no explicit comparison between the free and non-free-bank states will be attempted. One should bear in mind, however, that even if redemption rates in non-free-bank states were no higher than those in free bank states (and what evidence is contained in the auditors' reports suggests that they were lower in free bank states), the low failure rates in non-free-bank states imply that the losses to note holders were probably negligible.

Rolnick and Weber calculate a time-series for each of the four states that they call "the expected value of a randomly selected bank note held until 1863 as of the date of the condition report" (p. 1084). Actually, what they show are the *ex post* average 1863 values of the notes of banks that are operating in the given year. There are several problems with this as a measure of the riskiness of note holding. First, as with any *ex post* measure, it is based solely on information not available at the time that agents are making their decisions, and hence may bear little relation to the actual determinants of the notes' market values. Second, their calculations place very little structure on the data. Since it is a fairly general property of forecasts that they are less volatile than the time-series that they are attempting to predict, it seems likely that Rolnick and Weber's estimates are overly sensitive to the realizations on which they are based.

A more serious problem with their calculations is, once again, Rolnick and Weber's failure to take account of the truncation of their sample, one which leads them to make unjustified conclusions about the improvement in the performance of free banks over time. The basic problem is that the probability of a bank's failure *by 1863* will decrease over time, even with a constant hazard rate model as in the previous section. Thus Rolnick and Weber's measure will inevitably improve over time (and be biased upward to begin with). Yet they conclude from their calculations that, for example, "the safety of Minnesota free bank notes improved substantially after [July 1859]..." (p. 1087). Such a conclusion would validly follow from their results only if one were to assume that everyone knew for certain all along that after 1863 all bank notes would be safe forever.

TABLE 3—AVERAGE REDEMPTION RATES (R) AND HOLDING LOSSES (L)

	NY	WI	IN	MN
R	0.95	0.87	0.95	0.48
L	0.002	0.02	0.02	0.12

Note: Calculations are described in the text.

This section reports results from an alternative set of calculations. I calculate, for each of the four states, the average redemption rate (denoted R) of failed banks. The expected redemption rate for one year in the future is then just $1-p(1-R)$, where p is the failure probability. Put another way, the expected one-year holding loss (denoted L) is $p(1-R)$. Table 3 shows the results from this calculation, where p is the inverse of the life expectancy estimate from Table 1, and R is the average redemption rate for the state calculated from the Rolnick-Weber data.

Note that this procedure does not actually solve the problem of using information unavailable to agents. The statistics are based on the sample as a whole, and therefore make use of (relative to a particular year in the sample) both available and unavailable information. Ideally one would want to make "rolling" forecasts using just the information available in each year, but the data do not extend back in time far enough to do so. Thus I instead make an assumption about the stability of the process over the sample period. The procedure does get around the truncation problem, since the failure probabilities are consistent estimates (based on the procedures of the previous section), and the estimates of the average redemption rate, being conditioned on failure, are unaffected by the truncation of the sample. Although the results from Section I suggest that the quality of notes may not have been constant over time, one can interpret the results in Table 3 as measures of average quality over the period. While these calculations are less ambitious, and the results less "informative,"

they are not guilty of straining the limits of the data. What emerges is that the losses were not overwhelming (with the exception of Minnesota), but they were, except for the case of New York, nonetheless certainly non-negligible.

III. Conclusions

The preceding sections suggest that free banking legislation often resulted in very high failure rates in those states relative to failure rates in non-free-bank states. This was particularly so in the initial years following the enactment of the legislation, but remained true to some extent throughout the period. The high failure rates, combined with the frequent inability of failed banks to pay note holders the par value of their notes (as documented by Rolnick-Weber) meant that the holding of free bank notes was a moderately costly (and risky) activity. Thus I would caution against Rolnick and Weber's rather sanguine view of the free banking experiment.

On the other hand, it was certainly possible for free banking to be successful. New York's bank failure rate was not much higher than those of non-free-bank states, it experienced a large increase in the number of banks in the state, and the riskiness of its bank notes was negligible. Nonetheless, the experiences of other states indicate that such success may have been the exception rather than the rule. Wisconsin fostered growth in its banking industry, but had continuing difficulties with failures and losses to note holders. Indiana improved with regard to failures and losses, but ended up with many fewer banks than it had in the first year of the policy. Minnesota's problems with note-holder losses were especially severe. New Jersey probably had the best experience of the states examined in this paper other than New York, with most of its problems in the first few years. More importantly, the results presented herein show clearly that the rapid growth of a state's banking infrastructure did not require free banking laws. Maine, and Maryland to a lesser extent, experienced significant growth in the number of chartered banks, with less of the chaos that accompanied such growth in free banking states. The difficulty here is that such growth was subject to the whims of state legislators, who need not have been responsive to economic factors.

REFERENCES

Rolnick, Arthur J. and Warren E. Weber, "New Evidence on the Free Banking Era," *American Economic Review*, December 1983, *73*, pp. 1080–91.

U.S. Congress, *House Executive Documents*, "Report of the Secretary of the Treasury on the Condition of the State Banks," various years, 1837–63.

New Evidence on Free Banking in the United States

By HUGH ROCKOFF*

In their 1983 paper, Arthur Rolnick and Warren Weber report a number of important calculations concerning antebellum free banking in New York, Indiana, Wisconsin, and Minnesota. Their results clarify the dimensions of the free banking experiments and add further evidence that some portraits of the losses under antebellum free banking are exaggerated. But, while it is important to show that the actual losses were smaller than many have imagined, the losses and failure rates reported by Rolnick and Weber will not by themselves prove entirely persuasive. Skeptics will point to a high overall failure rate of 15 percent, and to failure rates of 25 to more than 50 percent in the western states, as evidence that free banking was unsuccessful. Part of the problem, I believe, is that Rolnick and Weber present these experiments as repeated trials of the same legislation. Here I show that if one looks carefully at the legal and historical circumstances in which each experiment took place, a case can be made that will satisfy even the skeptics who remain unconvinced by the losses and failure rates presented by Rolnick and Weber.

I illustrate this by referring first to the experience in an additional state, Michigan. Rolnick and Weber mention Michigan in a footnote (p. 1089), but their neglect of the particular circumstances attending these experiments leads them to an excessively optimistic view of what happened there. I believe the experience in Michigan is important because it influenced many of the traditional accounts. For example, the famous story about kegs of nails with coins on top being moved from bank to bank one step ahead of the bank examiners, that almost came to summarize antebellum free banking (which is included in the quote from Bray Hammond cited by Rolnick-Weber, p. 1083) is undoubtedly a reference to Michigan. (See Hammond, 1957, p. 601.)

Rolnick and Weber cite a Michigan bank commissioners' report (dated January 1839) to the effect that most of the free banks had sufficient assets to pay off at par, and go on to suggest that traditional pictures of the losses in Michigan are overdrawn. Previous generations of banking historians have been familiar with the commissioners' report but have interpreted it differently. Later material is hard to come by. One reason is that shortly after the commissioners' report was issued, Michigan abolished the bank commissioners' department, and transferred its duties to the attorney general, a fact that in itself tells us something about what was happening.

There is, however, an old paper by Alpheus Felch (1880) that gives a very different view from that suggested by Rolnick and Weber. Felch was writing long after the event, but there is good reason to respect his views. He was in the state legislature when the free banking law was passed, and was one of four to vote against it. He was, more importantly, one of the bank commissioners who authored the report cited by Rolnick and Weber. He was also a member of the Supreme Court of Michigan in 1844 when the free banking law was declared to be in conflict with the state constitution. According to Felch, the total amount of currency left outstanding by the free banks was not "ascertainable; but it could not in all probability be less than one million of dollars; and this fell as a dead loss on the community" (p. 123).[1]

The document that Felch seems to rely on the most is the *Attorney General's Report* dated December 1839 (Michigan, 1840). This

*Professor of Economics, Rutgers University, New Brunswick, NJ 08903.

[1]In my earlier study (1975, p. 17), I attempted to create a safe upper bound estimate of the total circulation of the free banks, but the result was probably too great an upper bound to be helpful. In my other work, I used Felch's figure (1974, p. 150).

document gives us a picture of Michigan's banking system eleven months after the bank commissioners' report. The *Report* lists 42 banks against which proceedings had begun for disolution of their charters (31 of them free banks), and bits of information on many of them provided by their receivers. The attorney general had some information on 34 of these banks on the basis of which he estimated their outstanding liabilities (mostly notes) at $1,663,033. In only 14 cases, however, was the receiver willing to guess the proportion of assets that were "good." For these 14 cases, good assets per dollar of liabilities was 23.6 cents; for the 12 free banks in this group it was 39.2 cents. But both figures may have been overly optimistic in the light of the on-going depression, and the eventual court decision that made further legal recourse impossible. If we blow up the attorney general's estimate for 34 banks to represent his total list of 42, and use the figure of 23.6 cents to represent what was eventually paid, then we are looking at losses on the order of $1.6 million. For the free banks alone, this reasoning would yield a lower figure, If we multiply the average liabilities for the free banks in the attorney general's list by 40, the number of free banks Felch thought got into operation, and assume they paid 39.2 cents on the dollar, we get total losses of almost precisely one million dollars, a number similar to Felch. These figures are still only approximate. But it is clear that they strongly support Felch, suggesting that until further evidence comes to light, Felch's venerable conclusion must stand as the most likely presumption about the fate of the Michigan free bank notes.

The failure of the banks to redeem at par, however, does not necessarily imply that the transfers from the public to the bankers approached this magnitude. Many of the notes may have entered circulation at a heavy discount—in Cincinnati, the notes of Michigan county banks were at a 50 percent discount by August 1838 and, by April 1839, Michigan notes could not be sold at all (Thomas Berry, 1943, p. 459). Later, some of the notes were bought by speculators at heavy discounts, who then brought suits against the bankers for the face value of the notes, so it is possible that in some cases there were net transfers from the bankers to the public. Felch's use of the term "dead loss" may be simply an assertion that the notes were not redeemed, rather than a reference to a net wealth transfer or loss in output, which is the construction one might be tempted to place on the phrase. In fact, it seems a reasonable conjecture that the willingness of the Michigan Supreme Court to declare the law unconstitutional, a decision that brought the process of suit and countersuit to a halt, was influenced by a realization that there would be little gain in fairness by further attempts to make the banks or their borrowers pay up.

The Michigan experience, moreover, was unique in several ways. The free banks were opened during a period of legal suspension of specie payments so that specie was necessary only because the legislation required a certain amount to be on hand. This was the circumstance that allowed the promoters to shift specie from bank to bank to keep one step ahead of the commissioners. In addition, the law permitted the banks to back up their notes with mortgages on the still largely undeveloped land of Michigan, an asset difficult to appraise, and of doubtful liquidity. Finally, it must not be forgotten that Michigan in 1840 was a frontier state, with a population slightly over 200,000; it had been admitted to the Union only a few years before the free banking experiment. The level of financial sophistication both in the general public and in the business community must have been fairly low. Later, in 1857, Michigan passed a second free banking law, and although the Civil War prevented a real test, there does not appear to have been any repetition of the boom-and-bust cycle. But circumstances had changed. For one thing, the second free banking law required the notes to be backed by government bonds —federal or certain selected states, including, of course, Michigan—and the law was written to force banks in most circumstances to back up each dollar in notes issued with more than a dollar's worth of securities.

The point is that to understand antebellum free banking it is necessary to examine

how the general idea was put into practice in each case. The central problem was how to insure confidence in the banking system while preserving the benefits of free entry and competition, the classic dilemma under any fractional reserve system. The mechanisms available for this purpose in the antebellum period were limited by institutional constraints such as the gold standard and by the vision of the legislators. The approach generally taken to reduce the risk of a panic, or restore confidence after one, was to tighten the secondary reserve requirement for notes. This, as we have seen, is what happened in Michigan. It also happened in Indiana, where the secondary reserve requirement was tightened in 1855 by requiring (to simplify a somewhat complex law) 110 percent in secondary reserves, when only 100 percent was required under the initial law, and by taking into account the lower yield on Indiana securities, (see my 1975 study, Table 12, p. 84). In New York the law was amended to exclude any bonds except those of New York. This greatly increased the security of the note issue since New York state bonds enjoyed a strong market independent of the banks, although the eligibility of mortgages remained a problem for a time.

Today, of course, we have alternative systems for preventing a loss of confidence—deposit insurance, and ultimately a central bank that can produce unlimited quantities of high-powered money and that stands ready to serve as a lender-of-last-resort. So the mechanism chosen by legislators in the antebellum period with its potential for distorting the investment policies of the banks is no longer necessary. The idea of a 100 percent reserve, however, which was the basic point, is not without merit. Although not now widely discussed, it was taken seriously as a method of stabilizing the banking system in the aftermath of the Great Depression, and was an important part of the Simons-Friedman tradition at Chicago. Indeed, Friedman (1959, pp. 75–76) could argue in the 1950's that the heavy investment in federal securities by the banking system had brought the system effectively halfway to a 100 percent reserve system. So antebellum free banking, although not always carried out successfully in detail, represented a sophisticated attempt to solve the problem of providing a competitive supply of banking services while maintaining the underlying stability of the system.

A second point illustrated by the Michigan example also applies more generally. It is one that I largely overlooked in my earlier work on antebellum free banking, and yet in retrospect seems obvious: the lack of financial sophistication in the communities struck by the worst experiences. Free banking experiments were less successful in Michigan, Indiana, and Minnesota than in Ohio, New York, and Louisiana for a variety of reasons. Part of the story must be that the former states were all frontier states where much of the business community had little experience in evaluating the potential demand for banking services, and where the public had little experience in evaluating the safety of alternative investments. The market tended to discourage the formation of unsound institutions, and eliminate them after they were formed, but it did not do it as smoothly as in established communities. Clearly, if we want to draw a lesson about the wisdom of free entry in modern circumstances, the experience of the developed states rather than the frontier states is what is relevant.

Finally, it is incumbent on the defender of antebellum free banking to show that free entry produced some of the benefits normally anticipated from it. Failure rates are only one dimension of the problem. I tried to do this in my earlier work, citing some comparative rates of return in banking, and related bits of evidence, suggesting that free entry improved the allocation of bank capital (1975, pp. 50–65). The best I could do, however, was to outline the case. Much remains to be done.

In short, I believe the results presented by Rolnick and Weber are important, but readers of their paper should not be left with the impression that the case for antebellum free banking rests solely on accepting their interpretation of the failure rates and related statistics. By showing that the means were at hand for increasing confidence in the system, and by showing that free entry produced a more efficient allocation of bank capital, a strong prima facie case can be made for

antebellum free banking even to someone who is distressed by the breakdowns that occurred in Michigan, Indiana, Minnesota, and certain other states.

REFERENCES

Berry, Thomas Senior, *Western Prices Before 1861: A Study of the Cincinnati Market*, Cambridge: Harvard University Press, 1943.

Felch, Alpheus, "Early Banks and Banking in Michigan," *Michigan Pioneer Collections 2*, 1880, 111–29.

Friedman, Milton, *A Program for Monetary Stability*, New York: Fordham University Press, 1959.

Hammond, Bray, *Banks and Politics in America: From the Revolution to the Civil War*, Princeton: Princeton University Press, 1957.

Rockoff, Hugh, "The Free Banking Era: A Reexamination," *Journal of Money, Credit and Banking*, May 1974, *6*, 141–67.

______, "The Free Banking Era: A Reexamination," in *Dissertations in American Economic History* (rev. doctoral dissertation, University of Chicago, 1972), New York: Arno Press, 1975.

Rolnick, Arthur J. and Weber, Warren E., "New Evidence on the Free Banking Era," *American Economic Review*, December 1983, *73*, 1080–91.

Michigan, *Attorney General's Report*, Senate Document 6, Annual Session, 1840.

Equilibrium Unemployment as a Worker Discipline Device: Comment

By Clive Bull*

In their very interesting and otherwise very clear paper, Carl Shapiro and Joseph Stiglitz (1984) make a claim concerning risk aversion and the lack of a private incentive to provide unemployment insurance that is misleading. This comment tries to clarify what can in fact be correctly claimed in this regard.

The technological basis of the Shapiro-Stiglitz model is a principal-agent model in which the firm, for which worker effort is a necessary input, cannot observe accurately the effort provided by the worker. Because of various restrictions that are placed on the employment contract that can be written, the only incentive that the firm can provide the worker in order to elicit effort is the threat of firing. Assuming that the probability of being hired by another firm in the period following being fired is less than one (as they show it will be in market equilibrium), then the seriousness of this threat is an increasing function of the difference between the wage paid when employed, w, and the wage paid when unemployed, w^u. Let there be a minimum wage, $\overline{w}$, that a firm must pay any worker that becomes unemployed after a separation from the firm, and for simplicity let $\overline{w}=0$. Shapiro and Stiglitz show (p. 437) that in a market equilibrium firms will set $w^u=\overline{w}=0$. The argument used is that if a firm were to raise w^u alone, it would reduce the incentive to shirk. In order to restore this incentive, it would then have to raise w. Thus given the constraint that the firm must elicit effort from the worker, its marginal rate of substitution between w^u and w is greater than one.[1] However, if the worker is risk neutral, his or her marginal rate of substitution between w^u and w is unity. Cost minimization by the firm then implies immediately that the optimal trade will involve $w^u=\overline{w}=0$.

In their basic model, workers are risk neutral. However, the authors consider three extensions of the model (p. 441)—one of which is to risk-averse workers. In that extension they claim that, in a market equilibrium, w^u will still equal $\overline{w}=0$ because no firm has an incentive to offer $\overline{w}>0$. This they rightly claim is surprising, as their model also includes exogenous random separations which occur with probability $b>0$ per period. Thus even a worker who always provided effort would become unemployed with probability one during his or her infinite lifetime. As saving (self-insurance) is ruled out in the model, there is a demand by workers for unemployment insurance, that is, $w^u>0$. To quote Shapiro and Stiglitz, "It is striking that the market provides no unemployment benefits even when workers are highly risk averse. Clearly the social optimum involves $\overline{w}>0$, if risk aversion is great enough. This may provide a justification for mandatory minimum benefit levels" (p. 441). In view of the implicit contract literature, this is indeed striking.

The claim that $w^u=\overline{w}$ no matter how great the risk aversion of workers and the implication Shapiro-Stiglitz tentatively draw for public policy are misleading as they rely implicitly on the assumption that w^u must be paid to unemployed workers after separation no matter how the separation occurred; that is, whether the separation was a random event or a firing because of shirking. However, the cause of separation does play a crucial role in the treatment of workers after separation in U.S. union contracts and in the labor laws of many OECD countries. Certainly in the model at hand there is a clear private incentive for a firm to offer insurance against the probability of a random sep-

*Department of Economics, New York University, 269 Mercer Street, New York, NY 10003. Research support from the National Science Foundation under grant no. SES-8409276 is gratefully acknowledged.

[1]As Shapiro-Stiglitz note in their reply, the use of the term marginal rate of substitution here is misleading. It refers to the slope of the incentive compatibility constraint that the firm faces.

aration just as in the implicit contract literature. Provision of this insurance would enable the firm to attract labor at a lower expected wage cost while leaving the disincentive to shirk untouched, though it will of course effect the incentive to find a new job once a random separation has occurred.

The Shapiro-Stiglitz claim that their model generates no private incentive to provide unemployment insurance is incorrect if we allow firms to make insurance conditional on the mode of separation. Such conditional insurance, of course, requires that third parties such as the courts or the NLRB have access to the firm's information concerning the worker's provision of effort or information concerning the exogenous events that cause separations other than fires. If such information is imperfect, then the firm will be tempted to breach the insurance contract by claiming to have fired for reasons of low effort workers who have in fact left the firm for other reasons. This type of unfair dismissal will naturally effect the optimal insurance contract. In view of this, among the many interesting questions that can be asked within the context of the Shapiro-Stiglitz model are, what is the impact of changing the ability of the courts to monitor the causes of separations on the equilibrium unemployment rate, and, if such monitoring is costly, what is the optimal level to be used?

REFERENCE

Shapiro, Carl and Stiglitz, Joseph E., "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, *74*, 433–44.

Equilibrium Unemployment as a Worker Descipline Device: Reply

By CARL SHAPIRO AND JOSEPH E. STIGLITZ*

We agree with Clive Bull's comment on two points, but take slight issue with him on a third. We agree that our 1984 paper was very interesting. We also agree that if workers are risk averse and if quits could easily be differentiated from fires, then the firm would offer unemployment insurance for workers who quit. Bull's remarks about why such insurance is not offered in our model, and regarding the effect of such insurance on equilibrium, are not quite correct, however.

Because of space limitations, in our earlier paper we were unable to expand on why we assumed that firms could not distinguish quits from fires for the purposes of unemployment compensation. Neither did we fully describe the role that this assumption played in our analysis. In fact, the equilibrium would be very different if employers could distinguish between quits and fires. If such distinctions are possible, and if workers were risk neutral, employers would shift much of workers' compensation into severance pay for quitters, yet provide no more than the legal minimum of pay ($\bar{w}$) to fired shirkers. This shift would reduce the equilibrium unemployment rate, possibly to zero. If workers are risk averse, however, such compensation shifting would be limited, and equilibrium unemployment would persist. In either case, Bull's remark that employers' ability to distinguish quits from fires would have "no impact on the incentive to shirk" is incorrect. In the new equilibrium we would need to distinguish V_Q, the expected lifetime utility of unemployed workers who had quit their jobs, and V_F, expected utility for those who were fired.

In our earlier paper we made the assumption that employers would find it rather difficult to distinguish between quits and fires. In particular, it would appear difficult to prove that a worker shirked, rather than quit, especially if the differences in payments to workers in those two groups is large. The reason we did not permit such distinctions was that it seemed to us, in general, to be a reasonably easy matter for either side of the contract to convert a "fire" into a "quit" and vice versa.

Assume that a worker receives unemployment compensation if he quits, but not if he is fired (in contrast with our public unemployment compensation programs). Then a worker who has shirked and been caught at it has an incentive to run to his boss and announce that he has quit, before he receives a notice that he is being fired. In practice, firms frequently give workers who have misbehaved a "second chance"; but if the worker then could announce that he was quitting, any such second chance would prove costly to the firm. Similarly, the firm, upon hearing rumors that, for one reason or another, one of its employees was contemplating quitting, would immediately fire him, claiming that he was shirking.

The costs associated with monitoring whether a worker had or had not shirked seem large; and the costs imposed on the employer, were he required to document thoroughly every firing, seem sufficiently great as to make this an unattractive policy. The considerations that lie behind this discussion are clearly beyond the scope of the formal model we developed; and clearly there are circumstances in which it is possible to differentiate between quits and fires. Such a distinction relies upon an outside (third) party observing workers' effort levels, however. Our analysis applies to situations (that are common) in which employers find it hard to observe effort; third parties will in general find effort monitoring at least as costly. In practice, monitoring and "management" or

*Woodrow Wilson School of Public and International Affairs, and Department of Economics, respectively, Princeton University, Princeton, NJ 08544.

"supervision" are joint products, and third-party monitoring is not observed. Any modeling requires simplification. In our judgment, the assumption that it was impossible to differentiate between quits and fires seemed better than the alternative polar assumption.

Bull's point about the importance of firms' abilities to distinguish quit and fires is legitimate. He seems, however, to have missed one of the two central messages of our paper. Whether or not one can or cannot distinguish between quits and fires, the equilibrium contract is likely not to be constrained Pareto efficient.

First, when there is full employment (as in most of the implicit contract literature), the incentive for providing unemployment insurance is that the firm can obtain workers at lower total expected (present discounted value of) costs. Here, the firm has no difficulty obtaining workers: there is a reserve army of the unemployed. Bull's remarks about the marginal rate of substitution (*MRS*) between wages, w, and unemployment compensation, w^u, reveal that he is missing this point: workers will trade off w and w^u (i.e., have a *negative MRS*), while firms find that a higher w^u requires a higher w; firms have a *positive MRS*. It is the differing *signs* of these *MRS*s, not the magnitudes, that is the fundamental problem. No amount of risk aversion will change this sign pattern. It is for this reason that firms will provide no more than the legal minimum of unemployment compensation if they cannot distinguish quits and fires. In general, there is no reason to expect the equilibrium rate of unemployment insurance, so determined, to be socially optimal. The socially optimal rate of unemployment benefits balances the insurance aspects of such benefits against the added monitoring costs they necessitate. Firms, however, do not account for the insurance aspects of unemployment benefits since they have no trouble attracting unemployed workers (especially when such benefits are low!).

Second, in the case where quits can be distinguished from fires, the decision to provide unemployment insurance by one firm imposes an externality on other firms. This externality arises because unemployment insurance influences the quit rate, which in turn affects the turnover rate. Increasing the turnover rate reduces the expected duration in the unemployment pool, and hence necessitates an increase in monitoring expenses and/or wages on the part of other firms to avoid shirking. This suggests that the firm may provide too high an unemployment benefit.

The exact balancing of the considerations that go into determining the socially optimal level of unemployment compensation is a complicated matter, which would take us beyond the scope of this reply. It is clear, however, that whether firms can or cannot differentiate between quits and fires, the provision of unemployment insurance by employers will not be at a socially optimal level.

REFERENCES

Bull, Clive, "Equilibrium Unemployment as a Worker Discipline Device: Comment," *American Economic Review*, September 1985, *75*, 890–91.

Shapiro, Carl and Stiglitz, Joseph E., "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, *74*, 433–44.

Capacity, Output, and Sequential Entry: Comment

By STANLEY S. REYNOLDS*

Two behavioral assumptions that are often made in the industrial organization literature are that an established firm (or group) may deter entry either through limit pricing (the Sylos Postulate) or by holding excess capacity (the Excess Capacity Hypothesis). In an interesting recent article in this *Review* (1981), Daniel Spulber examines these behavioral assumptions to see whether they are consistent with rational behavior by an established firm. Spulber's analysis is based on a two-firm, two-period game model in which the established firm is given a first-in advantage. By introducing this dynamic element into the model, Spulber is able to explicitly address the issue of the *optimality* of entry-deterring behavior.

Spulber finds that the use of limit pricing and/or excess capacity to deter entry is rational only under a very limited set of circumstances.[1] In particular, when the second-period outcome is determined by a Cournot-Nash equilibrium, he derives the following results. 1) The first-period output of the established firm is always less than or equal to the first-period output produced by a firm not anticipating entry. The established firm essentially accommodates entry and limit pricing does not occur. 2) The established firm never holds more capital than the amount that would minimize its production costs, given its output choices in periods one and two.

This comment takes issue with Spulber's conclusions about the Cournot-Nash case. It will be shown that the two results cited above may be reversed when the production technology is characterized by variable proportions. This reversal hinges on the particular type of Nash equilibrium employed in the analysis of the two-period model. Spulber implicitly uses a Nash equilibrium that is not subgame perfect.[2] It is shown below that, when one requires the Nash equilibrium to be subgame perfect, both limit pricing and excess capital investment outcomes are possible for the variable proportions technology case. The subgame perfection property thus seems to capture an important strategic element in decision making for the established firm. In some cases, this type of strategic behavior leads to entry barriers that would not exist under "innocent" profit maximization by the established firm. Strategic entry barriers are discussed by Steven Salop (1979). Spulber's notation and assumptions about demand and costs are adopted below.

I. A Subgame Perfect Nash Equilibrium

The period-two competition between the established firm and the potential entrant is a subgame for the two-period model, if the initial capital investment by the established firm may be observed by each firm prior to making its period-two decision. In this subgame, each firm chooses its output to maximize its period-two profit, given the capital stock carried over from period one by the established firm. The subgame equilibrium outputs for the established firm and potential entrant may be expressed as $x_2^{1*}(k^1)$ and $x^{2*}(k^1)$, respectively. The equilibrium outputs depend on the established firm's capital stock, k^1, because changes in k^1 alter the marginal cost of output for the established firm. Given Spulber's assumptions, there is a unique pair of subgame equilibrium outputs for each k^1. In addition, x^{2*} is continuous and nonincreasing in k^1 and x^{2*} is differen-

*Department of Economics, University of Arizona, Tucson, AZ 85721. I am grateful to Ron Braeutigam, Mark Isaac, Steve Matthews, and Dave Nickerson for helpful comments on an earlier draft.

[1] These conditions include Stackelberg leadership by the established firm in the post-entry game and the Stackelberg equilibrium output for the established firm must exceed its monopoly output.

[2] The concept of subgame perfection is discussed in Reinhard Selten (1975).

tiable with respect to k^1 except at the corner solution where x^{2*} falls to zero.

The marginal value of initial capital after entry (for an interior solution of the subgame) is composed of two terms (this corresponds to Spulber's equation (12), p. 508):

$$(1)\quad V^{N\prime}(k^1) = -c_k^1(x_2^1, k^1) + x_2^1 p'(x_2^1 + x^2)\frac{dx^{2*}}{dk^1}.$$

The second term on the right-hand side (which is positive for an interior solution) reflects the marginal value of capital for the established firm from reducing the entrant's output in period two. In a subgame perfect equilibrium, the established firm recognizes the impact of its capital choice on the period-two subgame and takes this into account when choosing its capital. In contrast, this second term is absent from Spulber's formulation. Spulber implicitly utilizes an open-loop Nash equilibrium concept for the two-period game.[3] In this case, the established firm chooses its initial capital given some fixed output choice by the potential entrant.

Given the marginal value of capital in (1), the first-order necessary conditions for the established firm's period-one choice of output and capital are (these correspond to Spulber's (13) and (14), p. 509):

$$(2)\quad 0 = p'(x_1^1)x_1^1 + p(x_1^1) - c_x^1(x_1^1, k^1),$$

$$(3)\quad 0 = -q - c_k^1(x_1^1, k^1) - \left(\frac{1}{1+r}\right)c_k^1(x_2^1, k^1) + \left(\frac{1}{1+r}\right)x_2^1 p'(x_2^1 + x^2)\frac{dx^{2*}}{dk^1}.$$

The final term on the right-hand side of (3) is positive for an interior solution in which entry occurs. The presence of this term has two implications. First, the established firm necessarily invests in more capital than the amount that would minimize the discounted present value of its total costs, given its output choices. This result corresponds to the excess investment result for the variable proportions case derived by Avinash Dixit (1980) for a one-period model. Second, the excess capital investment in a subgame perfect equilibrium may induce the established firm to produce more output in the first period than a monopolist not anticipating entry. So, a limit-pricing outcome with period-one price below the "monopoly price" is possible in a Nash equilibrium.

For the Cournot-Nash case, Spulber finds that entry is deterred only in Bain's case of "blockaded entry" (Proposition 3, p. 509); that is, entry is deterred in equilibrium only if the decisions made by a monopolist not anticipating entry would be sufficient to deter entry. This result does not hold for a subgame perfect Nash equilibrium. An entry deterrence equilibrium involves a corner solution in the period-two subgame. The marginal value of capital in period two for this case may be written as

$$(4)\quad V^{N\prime}(k^1)^- = -c_k^1(x_2^1, k^1) + x_2^1 p'(x_2^1)\frac{dx^{2*-}}{dk^1} > V^{N\prime}(k^1)^+ = -c_k^1(x_2^1, k^1)$$

where the minus and plus superscripts denote left- and right-hand derivatives, respectively. The first-order conditions for period-one capital choice are a pair of inequalities:

$$(5a)\quad 0 \le -q - c_k^1(x_1^1, k^1) + \left(\frac{1}{1+r}\right)V^{N\prime}(k^1)^-;$$

$$(5b)\quad 0 \ge -q - c_k^1(x_1^1, k^1) + \left(\frac{1}{1+r}\right)V^{N\prime}(k^1)^+.$$

An equilibrium for which the inequality in (5b) is strict represents a case of strategic equilibrium entry deterrence. That is, the established firm expands its period-one capital (and output) beyond the "monopoly" level in order to deter entry in period two. Such an equilibrium necessarily involves limit

[3]A definition of an open-loop Nash equilibrium is provided in Finn Kydland (1975). It is interesting to note that the open-loop Nash equilibrium analyzed by Spulber for the fixed proportions case is subgame perfect. The notion of subgame perfection becomes important in the variable proportions case because of the marginal-cost-reducing effect of increased capital investment.

pricing and excess capital investment. An example of this type of equilibrium is provided in an extended version of this paper (see my 1983 paper).

The results provided above indicate that the Sylos Postulate and excess capital investment need not be inconsistent with Nash equilibrium behavior. With a variable proportions technology, the strategic incentives for capital investment can lead an established firm to invest in more than the cost-minimizing quantity of capital and to set a preentry price below the "monopoly price."

REFERENCES

Dixit, Avinash, "The Role of Investment in Entry Deterrence," *Economic Journal*, March 1980, *90*, 95–106.

Kydland, Finn E., "Noncooperative and Dominant Player Solutions in Discrete Dynamic Games," *International Economic Review*, June 1975, *16*, 321–35.

Reynolds, Stanley, "Capital Investment, Limit Pricing and Entry," Discussion Paper, University of Arizona, September 1983.

Salop, Steven, "Strategic Entry Deterrence," *American Economic Review Proceedings*, May 1979, *69*, 335–38.

Selten, Reinhard, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 1975, *4*, 25–55.

Spulber, Daniel F., "Capacity, Output, and Sequential Entry," *American Economic Review*, June 1981, *71*, 503–14.

Capacity, Output, and Sequential Entry: Reply

By DANIEL F. SPULBER*

In his comment, Stanley Reynolds provides a very interesting application of the subgame perfect equilibrium concept to my 1981 entry model.[1] Reynolds claims that contrary to my analysis, "the Sylos Postulate and excess capital investment need not be inconsistent with Nash equilibrium behavior" (p. 896). It is not overly surprising if the characteristics of the equilibrium are altered by applying a different solution concept. However, the conclusions arrived at in my earlier paper regarding the Sylos Postulate and the Excess Capacity Hypothesis are quite robust to changes in the equilibrium. Reynold's assertion that the behavior of the incumbent at the subgame perfect equilibrium is consistent with the Sylos Postulate or the Excess Capacity Hypothesis is based on a misunderstanding of these concepts.

As Reynolds notes, the two-period, open-loop Nash equilibrium with capacity as an upper bound on output, which was examined in my earlier paper, is also a subgame perfect equilibrium. Thus, all of the results for this case still hold. In particular, the monopolist will deter entry only if the capacity level without the threat of entry exceeds the entry-deterring level. This requires capacity to be relatively inexpensive as compared to discounted marginal profit evaluated at the entry blocking output (Proposition 1 and equation (4) of my paper). Thus, "the Sylos Postulate is only satisfied in this limited sense" (p. 506). The established firm may choose, however, to permit entry. When entry occurs, the established firm *always* operates at full capacity before entry, thus contradicting the Excess Capacity Hypothesis. When capacity is relatively inexpensive, the established firm lowers its output to accommodate the entrant and holds excess capacity *after* entry, thus contradicting the Sylos Postulate.

For the case where capacity investment affects production costs, entrant and incumbent firm behavior at the subgame perfect equilibrium reinforces my conclusion that "the Sylos Postulate ignores both the strategic interaction between firms and the dynamic aspects of entry" (p. 503). The established firm at the subgame perfect equilibrium will not deter entry whether or not it is profitable to do so. Rather, entry will be deterred only if

$$(1)\quad q < -c_k^1(Q^1, k^1)\left(1+\frac{1}{1+r}\right) + \frac{1}{1+r}Q^1 p'(Q^1)\frac{dx^{2*}}{dk^1},$$

where Q^1 is the entry blocking output. This requires capacity to be inexpensive relative to the present value of cost savings and the discounted marginal revenue gain from entry deterrence. Otherwise, the established firm will accommodate entry by *lowering* its output, $x_1^1 > x_2^1$, at the subgame perfect equilibrium. This is hardly a confirmation of static limit pricing or the Sylos Postulate. Indeed, the threat of maintaining an unprofitably high constant output is precisely the type of *ad hoc* threat that would not be carried out by an established firm (or believed by a potential entrant) at a subgame perfect Nash equilibrium.

My conclusion that the Excess Capacity Hypothesis is "inconsistent with post-entry Cournot-Nash behavior whether or not entry is permitted by the established firm" (p. 504) is also reinforced by the results for the subgame perfect equilibrium. Consider first the case where entry is deterred. As Reynolds

*Department of Economics, University of Southern California, University Park, Los Angeles, CA 90089-0035. The support of the National Science Foundation is gratefully acknowledged. I thank Bob Becker and Jennifer Reinganum for helpful discussions. Any errors are my own responsibility.

[1]Additional extensions of my entry model are carried out for the case of uncertainty by S. Perrakis and G. Warskett (1983) and for a multiperiod perfect equilibrium by L. Arvan (1983), see also Reynolds (1983).

correctly points out, this *may* require investment (and output) in excess of the monopoly level. Output is kept *constant* at the pre-entry level, $x_1^1 = x_2^1 = Q^1$, and there is excess capacity in *both* periods, $q > -c_k^1(Q^1, k^1)(1 + (1/(1+r)))$.[2] However, entry is deterred by a large equilibrium output, not by the *threat* of raising output to capacity. The Excess Capacity Hypothesis asserts that it is the capacity level itself (and the threat to raise the low monopoly output) which deters entry. Again, this is precisely the type of *ad hoc* threat that would not be carried out by the established firm (or believed by the entrant) at the subgame perfect equilibrium.

When entry is not deterred, output falls, which certainly contradicts the Excess Capacity Hypothesis. Further, since $x_1^1 > x_2^1$ and $c_{xk} < 0, q > -c_k^1(x_2^1, k^1)(1+(1/(1+r)))$ which implies excess capacity in the second period for the established firm (as given by my equation (25)). The fact of excess capacity *after* entry certainly contradicts the Excess Capacity Hypothesis. Further, output and capacity may be lower than the monopoly level at the subgame perfect Nash equilibrium. This will hold unless capacity is sufficiently inexpensive (relative to cost savings) to overcome the lower returns to capacity in the post-entry equilibrium.

The appropriate game theoretic solution concept should be the one best suited to the economic problem at hand. The open-loop Cournot-Nash framework may best describe markets in which the entrant and established firm possess the same ability to contract forward. Thus, the firms compete for future market shares in the first period despite investment and production lags. There is evidence to suggest that this may be an accurate description of a number of industries, such as electronic data processing.[3] In any case, both the open- and closed-loop solutions yield insights into the the Sylos Postulate and the Excess Capacity Hypothesis.

[2] My earlier paper contains a misprint which does not affect the results as stated. The assumption on the cost function on p. 508 should read $c_{kk}^1 > 0$.

[3] Many firms manufacturing small business computers have an established consumer base which could be tapped for future sales of general purpose computer systems, (see Franklin Fisher et al., 1983, pp. 174–75). IBM has also faced entry from potential suppliers established in other fields (office equipment, components, and communications). Thus, contracts made in the current period for future delivery may accurately describe the entry challenge faced by established firms in this market.

REFERENCES

Arvan, L., "A Note on the Role of Investment in Entry Deterrence," Faculty Working Paper No. 985, University of Illinois, November 1983.

Fisher, Franklin M., McGowan, John J. and Greenwood, Joen E., *Folded, Spindled and Mutilated: Economic Analysis and U.S. v. IBM*, Cambridge: MIT Press, 1983.

Perrakis, S. and Warskett, G., "Capacity and Entry under Demand Uncertainty," *Review of Economic Studies*, July 1983, *50*, 495–512.

Reynolds, S. S., "Capacity, Output, and Sequential Entry: Comment," *American Economic Review*, September 1985, *75*, 894–96.

———, "Capital Investment, Limit Pricing and Entry," Discussion Paper, University of Arizona, September 1983.

Spulber, D. F., "Capacity, Output, and Sequential Entry," *American Economic Review*, June 1981, *71*, 503–14.

Preliminary Announcement of the Program

NINETY-EIGHTH ANNUAL MEETING
OF THE AMERICAN ECONOMIC ASSOCIATION

New York, New York, December 27–30, 1985

Friday, December 27, 1985

10:00 A.M. EXECUTIVE COMMITTEE MEETING

Saturday, December 28, 1985

8:00 A.M. EVALUATING THE NONPROFIT SECTOR
Presiding: RICHARD STEINBERG, Virginia Polytechnic Institute and State University
Papers: BURTON WEISBROD, University of Wisconsin-Madison
Informational Asymmetry and the Relative Performance of Nonproprietary Organizations
SUSAN ROSE-ACKERMAN, Columbia Law School
Nonprofit Firms: Are Government Grants Desirable?
MARK SCHLESINGER, Harvard University
Nonprofit Ownership and the Property Rights Model: A Reexamination
Discussants: GEORGE AKERLOF, University of California-Berkeley
JEFF STRNAD, School of Law, University of Southern California and California Polytechnic State University
KENNETH CLARKSON, School of Law, University of Miami

8:00 A.M. UNIONS IN DECLINE: CAUSES AND CONSEQUENCES
Presiding: RICHARD EDWARDS, University of Massachusetts-Amherst
Papers: RICHARD FREEMAN, Harvard University
The Effect of the Union Wage Differential on Management Opposition and Union Organizing Successes
RICHARD EDWARDS, MICHAEL PODGURSKY, AND PAUL SWAIM, University of Massachusetts-Amherst
Patterns of Recent Union Decline
MICHAEL WACHTER, University of Pennsylvania
What Problems are Unions Attempting to Solve?
Discussants: CHARLES CRAYPO, University of Notre Dame
MICHAEL REICH, University of California-Berkeley

8:00 A.M. POLITICS AND ECONOMIC POLICIES
Presiding: JACK E. ADAMS, University of Arkansas-Little Rock
Papers: JAMES E. ALT, Washington University-St. Louis
Party Strategies, World Demand, and Unemployment: The Political Economy of Economic Activity in Western Industrial Nations
K. ALEX CHRYSTAL, Sheffield University, England
What Can Economics Learn from Political Science and Vice Versa?
DOUGLAS A. HIBBS, JR., Harvard University
Political Parties and Macroeconomic Policies and Outcomes
HENRY W. CHAPPELL, JR., University of South Carolina, AND WILLIAM R. KEECH, University of North Carolina-Chapel Hill
Party Differences in Macroeconomic Policies and Outcomes
Discussants: LEWIS E. HILL, Texas Tech University
WALT MISIOLEK, University of Alabama
JACK E. ADAMS, University of Arkansas-Little Rock

8:00 A.M. SEX DIFFERENCES IN ACADEMIC AND NONACADEMIC LABOR MARKETS
Presiding: MARY FISH, University of Alabama
Papers: PAULA VOOS, University of Wisconsin-Madison
Wage Discrimination: A New Approach Based on the Direct Measurement of Productivity

DEBRA A. BARBEZAT, Amherst College
An Empirical Analysis of Salary Differentials by Sex: The Academic Labor Market
CLETUS C. COUGHLIN AND ANTHONY O'BRIEN, University of Georgia
Nonblind Refereeing as a Barrier to Entry in the Economics Profession
NADJA ZALOKAR, University of Florida
Generational Differences in Female Occupational Attainment—Have the 1970's Changed Women's Opportunities?
Discussants: BARRY T. HIRSCH, University of North Carolina-Greensboro
MARIANNE A. FERBER, University of Illinois-Urbana

8:00 A.M. GRADUATE STUDENT PAPER SESSION (Joint Session with Omicron Delta Epsilon)
Presiding: G. RANDOLPH RICE, Louisiana State University-Baton Rouge
(Papers and titles to be announced)

8:00 A.M. TOWARD A THEORY OF MINORITY BUSINESS DEVELOPMENT (Joint Session with the National Economic Association)
Presiding: GAVIN M. CHEN, U.S. Department of Commerce
Papers: GAVIN M. CHEN, U.S. Department of Commerce
Setting the Stage: A Theoretical and Historical Background
DAVID SWINTON AND JOHN HANDY, Clark College
The Theory
PETER BEARSE, Peter Bearse and Associates
Applied Theory
Discussants: TIMOTHY BATES, University of Vermont
BERNADETTE CHACHERE, Hampton Institute
SHERYL DOW-BAILEY, Old Dominion University
MARGARET SIMMS, The Urban Institute

8:00 A.M. ASIAN ECONOMIC STUDIES—JAPAN, CHINA, AND INDIA (Joint Session with the Committee on Asian Economic Studies)
Presiding: JAMES RIEDEL, The Johns Hopkins University
Papers: KAZUO SATO, Rutgers University
Increasing Returns and International Trade: The Case of Japan
JUSHAN BAI, Nankai University, TEH-WEI HU, Pennsylvania State University, AND SHUZHONG SHI, Nankai University (Tianjin, China)
Household Expenditure Patterns in a Large Chinese City
V. PANDIT, Delhi University and University of Pennsylvania
Growth Performance of the Indian Economy: Supply or Demand
Discussants: EDWARD LINCOLN, The Brookings Institution
NICHOLAS LARDY, University of Washington
LESLIE LIPSCHITZ, International Monetary Fund

10:15 A.M. AMERICAN ECONOMICS OFFICIALLY TURNS 100 (Joint Session with the History of Economics Society)
Presiding: A. W. COATS, Duke University and University of Nottingham
Papers: ROBERT M. SOLOW, Massachusetts Institute of Technology
What Do We Know that Francis Walker Didn't?
EARLENE CRAVER AND AXEL LEIJONHUFVUD, University of California-Los Angeles
Economics in America: What it Owes to the Intellectual Migration
EDWARD SHILS, University of Chicago and Cambridge University
A Sociologist's Perspective on the Economics Profession

10:15 A.M. ECONOMIC ISSUES IN IMMIGRATION POLICY
Presiding: THOMAS J. ESPENSHADE, The Urban Institute
Papers: BARRY R. CHISWICK, University of Illinois-Chicago Circle
The Employment and Employers of Illegal Aliens
WILFRED J. ETHIER, University of Pennsylvania
Illegal Immigration: The Host Country Problem
FRANCISCO L. RIVERA-BATIZ, Indiana University
Can Border Industries Be a Substitute for Immigration?
Discussants: JEAN BALDWIN GROSSMAN, Mathematica Policy Research, Inc.
FRANK DE LEEUW, U.S. Department of Commerce
MICHAEL PIORE, Massachusetts Institute of Technology

10:15 A.M. THE MARKET FOR CORPORATE CONTROL (Joint Session with the Association of Managerial Economists)
Presiding: HENRY G. MANNE, Emory University
Papers: HAROLD DEMSETZ, University of California-Los Angeles
The Control Function of Insider Trading
MARK HIRSCHEY, University of Colorado-Denver
Mergers, Buyouts, and Fakeouts
MICHAEL C. JENSEN, Harvard University and University of Rochester
The Market for Corporate Control
Discussants: JOHN J. MCCONNELL, Purdue University
DAVID W. MULLINS, JR., Harvard University

10:15 A.M. THE POLITICAL ECONOMY OF OUTER SPACE
Presiding: PAUL R. PORTNEY, Resources for the Future
Papers: LINDA COHEN, The Brookings Institution and University of Washington, AND ROGER NOLL, Stanford University
Commercializing Space
JOHN O. LEDYARD, Northwestern University
Incentive Compatible Space Station Prices
MOLLY K. MACAULEY, Resources for the Future
Regulation and Technical Change in Communications Satellites
Discussants: JOEL SCHERAGA, Rutgers University
MARCELLUS SNOW, University of Hawaii

10:15 A.M. ECONOMIC ISSUES IN U.S. INFRASTRUCTURE INVESTMENT
Presiding: DAMIAN J. KULASH, National Research Council
Papers: THEODORE KEELER, University of California-Berkeley
Infrastructure and Productivity: The Case of Trucking
GEORGE PETERSON, The Urban Institute
Effects of Different Finance Choices on Infrastructure Expenditure
KENNETH SMALL, University of California-Irvine, AND CLIFFORD WINSTON, The Brookings Institution
Welfare Effects of Optimal Investment in Highway Infrastructure
Discussants: JOSE A. GOMEZ-IBANEZ, Harvard University
GREGORY K. INGRAM, World Bank

10:15 A.M. COMPUTABLE GENERAL EQUILIBRIUM AND WORLD ECONOMIC DEVELOPMENT (Joint Session with the Society for Policy Modeling)
Presiding: HERBERT SCARF, Yale University
Papers: IRMA ADELMAN, University of California-Berkeley, ALEXANDRA SARRIS, University of Athens, Greece, AND DAVID WELLS, University of San Francisco
Planning under Uncertainty with *CGE* Models
SHERMAN ROBINSON AND LAURA TYSON, University of California-Berkeley
CGE Models of Socialist Countries
PETER DIXON AND G. A. MEAGHER, Melbourne University, Australia
Analyzing the Effects of Changes in the Structure of Taxation via the ORANI Model of the Australian Economy
LARRY GOULDER, Harvard University
Long-Run Dynamics on *CGE* Models
Discussants: ALFREDO PEREIRA, Stanford University
MICHAEL MONTIAS, Yale University
STEVEN S. CHANG, St. John's University
VICTOR GINSBERG, CORE, Brussels

10:15 A.M. AGRICULTURAL ECONOMICS: A REFLECTION ON CONTRIBUTIONS AFTER SEVENTY-FIVE YEARS (Joint Session with the American Agricultural Economics Associations)
Presiding: WILLIAM H. MEYERS, Iowa State University
Papers: JAMES P. HOUCK, University of Minnesota
Contributions to Economic Thought
CARL A. FOX, Iowa State University
Contributions to Quantitative Methods
S. R. JOHNSON, Iowa State University
Future Challenges for Agricultural Economics
Discussants: D. GALE JOHNSON, University Of Chicago
MARC NERLOVE, University of Pennsylvania

10:15 A.M. ECONOMIC THEORY AND THE STANDARD OF LIVING
Presiding: TERESA GHILARDUCCI, University of Notre Dame
Papers: JAMES DUESENBERRY, Harvard University
Consumption Theory and Its Implications for Evaluating Relative Well-Being
CLAIR BROWN, University of California-Berkeley
The Theory and Measurement of the American Standard of Living Since 1920
JULIET SCHOR, Harvard University
Evaluating the Contributions to Well-Being of Work vs. Leisure
Discussants: JOAN HANNON, University of California-Berkeley
TIBOR SCITOVSKY, Stanford University

10:15 A.M. SOME INTERTEMPORAL ISSUES IN MACROECONOMICS
Presiding: BRUCE SMITH, Federal Reserve Bank of Minneapolis
Papers: JEFFREY A. MIRON, University of Michigan
Seasonal Fluctuations and the Life Cycle–Permanent Income Model of Consumption
BENJAMIN BENTAL AND DEVORAH KANTOROWICZ, Technion, Israel
A Welfare Analysis of Internal Debt Repudiation
OWEN EVANS, International Monetary Fund
Determination of Preference Parameters from Aggregate Consumption Equations—A Critique
Discussants: LAWRENCE CHRISTIANO, Carnegie-Mellon University
MARTIN EICHENBAUM, Carnegie-Mellon University
MARK GERTLER, University of Wisconsin

10:15 A.M. REDISTRIBUTION ASPECTS OF INFLATION AND DISINFLATION—A ROUNDTABLE (Joint Session with the Association for the Study of the Grants Economy)
Presiding: WALTER ADAMS, Michigan State University
Panel: KENNETH E. BOULDING, George Mason University and University of Colorado
MANCUR OLSON, University of Maryland
THOMAS F. WILSON, American Fletcher National Bank
MARTIN SPECHLER, Harvard University and Indiana University
RICHARD D. BARTEL, *Challenge* Magazine
JANOS HORVATH, Butler University
ALLEN A. SCHMIDT, Michigan State University

10:15 A.M. UNDERSTANDING THE FED
Presiding: JAMES ANNABLE, First National Bank of Chicago
Papers: THOMAS MAYER, University of California-Davis
The Debate about Monetarist Policy Recommendations
THOMAS HAVRILESKY, Duke University
A Theory of the Current Monetary Regime
PAUL SPINDT, Board of Governors of the Federal Reserve System, AND VEFA TARHAN, Loyola University of Chicago
The Federal Reserve's New Operating Procedures: A Postmortem
GIKAS HARDOUVELIS, Barnard College
Economic News, Exchange Rates, and Interest Rates
Discussants: DAVID LINDSEY, Board of Governors of the Federal Reserve System
RAY LOMBRA, Pennsylvania State University
DANA JOHNSON, Northern Trust Company

10:15 A.M. TRADE PATTERNS AND PROTECTIONISM
Presiding: E. HAZEL DENTON, World Bank
Papers: ANITA M. BENVIGNATI, Federal Trade Commission
The Commodity Composition of U.S. Intra-Firm Exports
LAWRENCE P. BRUNNER, Central Michigan University
The Effect of Trade Restrictions on the U.S. Economy in the Great Depression
JAMES H. CASSING, University of Pittsburgh, AND ARYE L. HILLMAN, Bar-Ilan University, Israel
Graduation, Trade Preferences, and International Equity
GERNOT KLEPPER, Kiel Institute of World Economics, FRG
Endogenous Tariff Formation in a Political Market for Protection
Discussants: J. MICHAEL FINGER, World Bank
C. MICHAEL AHO, Council on Foreign Relations

10:15 A.M. SOUTH AFRICA AND THE UNITED STATES: IS THERE ECONOMIC INTERDEPENDENCE? (Joint Session with the National Economic Association)
Presiding: GAVIN M. CHEN, U.S. Department of Commerce

Papers: FRANKLIN THOMAS, Ford Foundation
Corporate America and U.S. South Africa Policy
FRED CURTIS, Drew University
The Political Economy of Labor Migration in Southern Africa
JAMES STEWART, Pennsylvania State University
The Economic Impact of U.S. Anti-Apartheid Legislation
GAVIN M. CHEN, U.S. Department of Commerce, AND STEPHANIE WILSON, Abt Associates
Aspects of the South African Infrastructure
Discussants: RONALD M. JOHNSON, International Monetary Fund
BARBARA JONES, Clark College
JULIANNE MALVEAUX, San Francisco State University

10:15 A.M. U.S. BISHOPS PASTORAL ON CATHOLIC SOCIAL TEACHING AND THE U.S. ECONOMY (Joint Session with the Industrial Relations Research Association)
Presiding: GEORGE G. HIGGINS, Catholic University of America
Panel: CHARLES WILBER, University of Notre Dame
ANTHONY DOWNS, The Brookings Institution
SHELDON FREEDMAN, United Automobile Workers
MICHAEL NOVAK, American Enterprise Institute
MICHAEL JOYCE, John M. Olin Foundation

10:15 A.M. DISTRIBUTIONAL EFFECTS OF LARGE INCREASES IN PHYSICIAN SUPPLY (Joint Session with the Health Economics Research Organization)
Presiding: DONALD E. YETT, University of Southern California
Papers: KATHRYN M. LANGWELL, Mathematica Policy Research, Inc., AND JOHN DRABEK, Bureau of Health Professions
Location Choices in a Market with Expanding Supply: The Effect of Community and Personal Characteristics on Career Decisions of 1974–78 Medical School Graduates
LOUIS P. GARRISON, JR., GERARD ANDERSON, AND GAIL R. WILENSKY, Project Hope
Impact of Increasing Physician Supply on Specialty Mix in Market Areas
JOHN DRABEK, PAUL SEDER, U.S. Department of Health and Human Services, AND JAMES THOMPSON, National Institute of Mental Health
Factors Influencing the Geographic Distribution of Mental Health Care Professionals
Discussants: JOSEPH P. NEWHOUSE, The Rand Corporation
BARBARA KEHRER, Henry J. Kaiser Family Foundation
RICHARD M. SCHEFFLER, University of California-Berkeley

12:30 P.M. AEA/AFA JOINT LUNCHEON
Presiding: ROBERT C. MERTON, Massachusetts Institute of Technology
Speaker: ANTHONY M. SOLOMON, Former President of the Federal Reserve Bank of New York

2:30 P.M. SYMPOSIUM ON THE SHARE ECONOMY
Presiding: ROBERT M. SOLOW, Massachusetts Institute of Technology
Panel: ROBERT E. HALL, Stanford University
SHERWIN ROSEN, University of Chicago
LAWRENCE H. SUMMERS, Harvard University
ALAN S. BLINDER, Princeton University
MARTIN WEITZMAN, Massachusetts Institute of Technology

2:30 P.M. 1985 DISTINGUISHED LECTURE ON ECONOMICS IN GOVERNMENT (Joint Session with the Society for Government Economists)
Presiding: DOUGLAS L. ADKINS, U.S. Congress, Office of Technology Assessment
Speaker: GEORGE JASZI, Former Director, Bureau of Economic Analysis
An Economic Accountant's Audit

2:30 P.M. OCCUPATIONS AND LABOR MARKETS: A CRITICAL EVALUATION
Presiding: ISABEL SAWHILL, The Urban Institute
Papers: WILLIAM T. BIELBY, University of California-Santa Barbara, AND JAMES N. BARON, Stanford University
Segregation Within Occupations
PETER DOERINGER, Boston University
Non-Competing Groups and Labor Market Segmentation
MARIANNE A. FERBER, University of Illinois-Urbana, CAROLE A. GREEN, University of South Florida, AND JOE L. SPAETH, University of Illinois-Urbana
Work Characteristics versus Occupations as Determinants of Earnings

Discussants: HEIDI I. HARTMANN, National Academy of Sciences
GLEN C. CAIN, University of Wisconsin

2:30 P.M. REGIONAL GROWTH PATTERNS: TRENDS, PROSPECTS, AND POLICY IMPLICATIONS
Presiding: ROGER BOLTON, Williams College
Papers: BENJAMIN CHINITZ, University of Lowell
The Regional Transformation of the American Economy in the Twentieth Century
BEN STEVENS, Regional Science Research Institute, AND GEORGE TREYZ, University of Massachusetts-Amherst
Future Prospects
CHARLES LEVEN, Washington University
Welfare and Policy Implications of Regional Decline
Discussants: JOSEPH TUREK, State University of New York-Albany
BENNETT HARRISON, Massachusetts Institute of Technology
GERALD A. CARLINO, Federal Reserve Bank of Philadelphia

2:30 P.M. OLIGOPOLISTIC MARKETS WITH PRICE-SETTING FIRMS
Presiding: BETH ALLEN, University of Pennsylvania
Papers: ERIC MASKIN, Harvard University
Existence of Equilibrium with Price-Setting Firms
BETH ALLEN, University of Pennsylvania, AND MARTIN HELLWIG, University of Bonn, FRG
Price-Setting Firms and the Oligopolistic Foundations of Perfect Competition
JOHN SUTTON, London School of Economics
Price Competition among Vertically Differentiated Products
Discussant: RICHARD SCHMALENSEE, Massachusetts Institute of Technology

2:30 P.M. THE MONETARY-FISCAL POLICY MIX: IMPLICATIONS FOR MACROECONOMIC PERFORMANCE
Presiding: ANDREW F. BRIMMER, Brimmer & Company, Inc.
Papers: ANDREW F. BRIMMER, Brimmer & Company, Inc., AND ALLEN SINAI, Shearson Lehman Bros. and New York University
The Monetary-Fiscal Mix: Short-Run Implications
WILLIAM J. BEEMAN AND FREDERICK C. RIBE, U.S. Congressional Budget Office
Monetary-Fiscal Mix and Long-Run Growth in an Open Economy
JAMES TOBIN, Yale University
The Monetary-Fiscal Mix: Long-Run Implications
Discussants: WILLIAM POOLE, Brown University
WALTER W. HELLER, University of Minnesota
ALICIA H. MUNNELL, Federal Reserve Bank of Boston

2:30 P.M. TRADE RESTRICTIONS AND THE STEEL INDUSTRY
Presiding: WASSILY W. LEONTIEF, New York University
Papers: JAMES T. H. TSAO, U.S. International Trade Commission and George Mason University
The Economic Effects of Trade Restrictions on U.S. Steel Mill Products
WILLIAM R. CLINE, Institute for International Economics
Protection and Adjustment in the International Steel Industry
ROBERT W. CRANDALL, The Brookings Institution
Trade Restrictions as a Barrier to Structural Adjustment in the U.S. Steel Industry
Discussants: HENDRIK S. HOUTHAKKER, Harvard University
JOHN W. SUOMELA, U.S. International Trade Commission

2:30 P.M. FINANCIAL INNOVATION: THE POLICY PRESCRIPTION
Presiding: ELINOR H. SOLOMON, George Washington University
Papers: ALLEN N. BERGER, GERALD A. HANWECK, AND DAVID B. HUMPHREY, Board of Governors of the Federal Reserve System
Economics of Scope and Scale in Banking
JOHN H. HOTSON, University of Waterloo
The Case for Financial Reform: Or Was Irving Fisher Right After All?
ROBERT E. LITAN, The Brookings Institution
The Risks of Financial Product Deregulation
Discussants: DWIGHT M. JAFFEE, Princeton University
ROBERT A. EISENBEIS, University of North Carolina
JAMES R. BARTH, George Washington University
RICHARD W. KOPCKE, Federal Reserve Bank of Boston

2:30 P.M. NEW RESEARCH DIRECTIONS ON SOCIALIST ECONOMIC SYSTEMS
Presiding: ED A. HEWETT, The Brookings Institution
Papers: JOHN P. BURKETT, University of Rhode Island
Slack and Shortage in Socialist Economies
JOHN B. HALL, Drexel University
Toward a Theory of Plan Bargaining in Hungary
SERGIO G. ROCA, Adelphi University
Management of State Enterprises: A Comparison of the Soviet Union and Cuba
Discussants: FREDERIC L. PRYOR, Swarthmore College
JUDITH THORNTON, University of Washington

2:30 P.M. CONFLICT AND PEACE ECONOMICS I: STRATEGY IN THE INTERNATIONAL SECTOR (Joint Session with the Peace Science Society (International))
Presiding: WALTER ISARD, Cornell University
Papers: MARTIN SHUBIK, Yale University
The Role of Game-Theoretic Analysis in International Conflict Research
BARRY NALEBUFF, Harvard University
Strategic Economic Factors in National Security
Discussants: STEVEN BRAMS, New York University
MANCUR OLSON, University of Maryland

2:30 P.M. RISK AND MANAGERIAL BEHAVIOR
Presiding: F. M. SCHERER, Swarthmore College
Papers: RICHARD D. MACMINN, University of Texas-Austin
Insurance and Corporate Risk Management
FREDERICK HARRIS, University of Texas-Arlington
Competing Theories of Firm Decision Making under Risk: An Empirical Test
DARYL N. WINN, University of Colorado-Boulder
Tobin's Q Ratio, Investment, and Executive Behavior: An Empirical Examination
JOHN LUNN, Louisiana State University-Baton Rouge
The Diffusion of New Technology: The Case of Pioneering Adopters
Discussants: STEPHEN ROSS, Yale University
WILLIAM MCEACHERN, University of Connecticut

2:30 P.M. WELFARE REFORM: NEW RESEARCH AND POLICY DEVELOPMENTS
Presiding: ROBERT MOFFITT, Brown University
Papers: ROBERT MOFFITT, Brown University
Work Incentives in the AFDC System: An Analysis of the 1981 Reforms
JUDY GUERON, DANIEL FRIEDLANDER, BARBARA GOLDMAN, AND DAVID LONG, Manpower Demonstration Research Corporation
Interim Findings from the Demonstration of State Work/Welfare Initiatives
THOMAS FRAKER, BARBARA DEVANEY, AND EDWARD CAVIN, Mathematica Policy Research, Inc.
An Evaluation of the Effect of Cashing Out Food Stamps on Food Expenditures
Discussants: HENRY AARON, The Brookings Institution and University of Maryland
HAROLD WATTS, Columbia University
EDWARD GRAMLICH, University of Michigan

2:30 P.M. OVERSHOOTING AGRICULTURAL COMMODITY MARKETS AND PUBLIC POLICY (Joint Session with the American Agricultural Economics Association)
Presiding: G. EDWARD SCHUH, World Bank
Papers: GORDON C. RAUSSER, JAMES A. CHALFANT, H. ALAN LOVE, YASUO NISHIYAMA, AND KOSTAS STAMOULIS, University of California-Berkeley
Macroeconomic Linkages, Taxes, and Subsidies on the U.S. Agricultural Sector
MARGARET ANDREWS, Rutgers University
Taxes/Subsidies and the Political Economy of Agricultural Policies
Discussants: JEFFREY A. FRANKEL, University of California-Berkeley
MAURICE OMSTFELD, Columbia University

4:30 P.M. TAX POLICY FOR ECONOMIC GROWTH
Presiding: JOSEPH A. PECHMAN, The Brookings Institution
Panel: MARTIN S. FELDSTEIN, Harvard University and National Bureau of Economic Research
ALAN GREENSPAN, Townsend-Greenspan & Co., Inc.
WALTER W. HELLER, University of Minnesota
WILLIAM D. NORDHAUS, Yale University

8:00 P.M. RICHARD T. ELY LECTURE
Presiding: ALICE M. RIVLIN, The Brookings Institution
Paper: HERBERT STEIN, American Enterprise Institute
(Title to be announced)

Sunday, December 29, 1985

8:00 A.M. FEDERAL DEFICIT AND THE GRANTS ECONOMY (Joint Session with the Association for the Study of the Grants Economy)
Presiding: KENNETH E. BOULDING, George Mason University and University of Colorado
Papers: JOSEPH SCHERER, Economic Consultant, Georgetown, Maine
The Grant Implications of Federal Tax and Spending Reforms
THOMAS MULLER, The Urban Institute
The Spatial Redistribution Effect of Some Federal Budget Reform Scenarios
(To be announced)
Measuring the Size and Leverage of Subsidy Components in Off-Budget Federal Financing
Discussant: HOWARD P. TUCKMAN, Memphis State University

8:00 A.M. ROUNDTABLE ON ECONOMIC EDUCATION: INCREASING THE PUBLIC'S UNDERSTANDING OF ECONOMICS
Presiding: MARIANNE A. FERBER, University of Illinois-Urbana
Papers: ALBERT E. REES, Alfred P. Sloan Foundation
The Marketplace of Economic Ideas
LEONARD M. SILK, *The New York Times*
Communicating Economic Ideas and Controversies
MICHAEL A. MACDOWELL, Joint Council on Economic Education
What Can We Expect of the Schools?
W. LEE HANSEN, University of Wisconsin-Madison
Building a Knowledge of Economics among College Graduates

8:00 A.M. *LDC* POLICY RESPONSES TO EXOGENOUS SHOCKS
Presiding: RONALD I. MCKINNON, Stanford University
Papers: BELA BALASSA, The Johns Hopkins University and World Bank
LDC Policy Responses to Exogenous Shocks
MAXWELL J. FRY, University of California-Irvine
LDC Monetary Policy Responses to Exogenous Shocks
MOHSIN S. KHAN, International Monetary Fund
LDC Exchange Rate Policy Responses to Exogenous Shocks
VITO TANZI, International Monetary Fund
LDC Fiscal Policy Responses to Exogenous Shocks
Discussants: ALBERTO GIOVANNINI, Columbia University
NATHANIEL LEFF, Columbia University

8:00 A.M. GENDER AND RACE IN THE ECONOMICS CURRICULUM
Presiding: LESTER THUROW, Massachusetts Institute of Technology
Papers: ROBIN BARTLETT, Denison University
Introducing the Concepts of Race, Gender, and Class into the Principles Curriculum
BERNADETTE CHACHERE, Hampton University
Reflections on the Treatment of Race in the Economics Curriculum
SUSAN FEINER, Virginia Commonwealth University, AND BARBARA MORGAN, London School of Economics
Hidden by the Invisible Hand: Race and Gender in Introductory Economics Textbooks
Discussant: RICHARD WOLFF, University of Massachusetts-Amherst

8:00 A.M. THE EFFICIENT PROVISION OF LOCAL PUBLIC SERVICES
Presiding: ROBERT HARTMAN, U.S. Congressional Budget Office
Papers: ANWAR M. CHAUDRY-SHAH, Department of Finance, Canada
Capitalization and the Theory of Local Public Finance: An Interpretative Essay
RANDALL CRANE, University of Wisconsin-Milwaukee
Efficient Policies for Local Economies
STEVEN BENDER, York University
Revealing Preferences for Public Goods through Referenda
DAVID E. WILDASIN, Indiana University
Tax Exporting and the Marginal Cost of Public Funds

Discussants: JOHN YINGER, Harvard University
ANDREW RESCHOVSKY, Tufts University
KATHARINE BRADBURY, Federal Reserve Bank of Boston
DENNIS ZIMMERMAN, Congressional Research Service

8:00 A.M. WAGES, UNEMPLOYMENT, AND INVOLUNTARY RETIREMENT
Presiding: JOSEPH G. ALTONJI, Columbia University
Papers: ESKANDER ALVI, University of Arizona
Quasi-Fixity and the Uneven Incidence of Layoffs in a Search Equilibrium
JONATHAN MICHIE, Oxford University, England
The Cyclical Behavior of Wages
FUSUN F. GONUL, Ohio State University
A Methodology for Determining Whether Unemployment and Out of the Labor Force are Distinct States
JAGDISH HANDA, McGill University
A Theory of Mandatory Retirement
Discussants: SHELLY LUNDBERG, University of Washington
LARRY KATZ, University of California-Berkeley
LISA LYNCH, Massachusetts Institute of Technology

8:00 A.M. NEW INSIGHTS ON THE STRUCTURAL ECONOMIC POSITION OF AFRICAN-AMERICANS (Joint Session with the National Economic Association)
Presiding: CLEVELAND CHANDLER, Howard University
Papers: FRANK G. DAVIS, Howard University
A New Approach to Urban Poverty: The Case of the D.C. Economy
LORENZO BROWN, Federal Trade Commission
A Comparative Analysis of Industrial and Market Structure among African-Americans and in the U.S. Economy as a Whole
JAMES HEFNER, Jackson State University
The Sunbelt and the Snowbelt: Industrial and Market Structure among African-Americans and the U.S. Economy
CLEVELAND CHANDLER, Howard University
The Structure of the Economic Base among African-Americans in the U.S. Economy
Discussants: WILLIAM DARITY, JR., University of North Carolina-Chapel Hill
RHONDA M. WILLIAMS, University of Texas-Austin

8:00 A.M. FIRST FORAYS INTO THE HISTORY OF ECONOMICS (Joint Session with the History of Economics Society)
Presiding: DAVID LEVY, George Mason University
Papers: ROY EPSTEIN, University of Illinois-Chicago Circle
Cowles' Problems Revisited
MARY MORGAN, University of York, England
Statistics Without Probability and Haavelmo's Revolution
CHRISTOPHER GILBERT, Oxford University, England
The Development of the British School in Econometrics
Discussant: EDWARD E. LEAMER, University of California-Los Angeles

8:00 A.M. THE SHIFTING PATTERNS OF COMPARATIVE ADVANTAGES—UNITED STATES–ASIA (Joint Session with the Committee on Asian Economic Studies)
Presiding: RICHARD HOOLEY, University of Pittsburgh
Panel: ROBERT E. BALDWIN, University of Wisconsin
LAWRENCE B. KRAUSE, The Brookings Institution
RANDOLPH BARKER, Cornell University
PAUL P. STREETEN, Boston University and World Bank

8:00 A.M. AGGLOMERATION ECONOMIES: THEORY AND MEASUREMENT (Joint Session with the American Real Estate and Urban Economics Association)
Presiding: EDWIN S. MILLS, Princeton University
Papers: ROBERT W. HELSLEY, University of British Columbia
Agglomeration and External Economies
GERALD A. CARLINO, Federal Reserve Bank of Philadelphia, AND EDWIN S. MILLS, Princeton University
Do Urban Areas Matter?
RONALD L. MOOMAW, Oklahoma State University
Urbanization and Localization Economies: Estimates at the 2- and 3-Digit Levels

Discussants: MASHISA FUJITA, University of Pennsylvania
MICHAEL S. FOGARTY, Federal Reserve Bank of Cleveland
RANDALL W. EBERTS, University of Oregon

8:00 A.M. CURRENT ISSUES IN MONEY, CREDIT, AND BANKING (Joint Session with the *Journal of Money, Credit and Banking*)
Presiding: STEPHEN MCCAFFERTY AND J. HUSTON MCCULLOCH, Ohio State University
Papers: GREGORY HOELSCHER, Chase Manhattan Bank
New Evidence on Deficits and Interest Rates
MADELYN ANTONCIC, Federal Reserve Bank of New York
High and Volatile Real Interest Rates: Where Does the Fed Fit In?
ANDREW K. ROSE, Massachusetts Institute of Technology
Money Demand in the 1980's

8:00 A.M. ADVANCES AND PROBLEMS IN VALUING ENVIRONMENTAL RESOURCES (Joint Session with the Association of Environmental and Resource Economics)
Presiding: WILLIAM D. SCHULZE, University of Colorado
Papers: NANCY BOCKSTAEL, University of Maryland
Issues in the Application of the Travel Cost Method for Valuing Environmental Amenities
ROBERT MENDELSOHN, Yale University
Identification, Simultaneous Equations, and Collinearity Problems in the Estimation of Hedonic Gradients
DON COURSEY, University of Wyoming
Continent Valuation of Nonmarket Goods: Lessons from the Laboratory
Discussants: V. KERRY SMITH, Vanderbilt University
MAUREEN CROPPER, University of Maryland
ROBERT ROWE, Energy and Resource Consultants

10:15 A.M. CHANGES IN WAGE NORMS
Presiding: CLAIR BROWN, University of California-Berkeley
Papers: ASSAR LINDBECK, Institute of International Economic Studies, Sweden, AND DENNIS SNOWER, Birkbeck College, England
Wage Setting, Unemployment, and Insider-Outsider Relations
MICHAEL L. WACHTER, University of Pennsylvania
Sources of Union Wage Rigidity
GEORGE L. PERRY, The Brookings Institution
Shifting Wage Norms and their Implications
DANIEL J. B. MITCHELL, University of California-Los Angeles
Unions vs. Nonunion Wage Norm Shifts

10:15 A.M. DO GOVERNMENT PROGRAMS CLOSE THE RACIAL GAP? (Joint Session with the National Economic Association)
Presiding: LYNN C. BURBRIDGE, The Urban Institute
Papers: EMMETT CARSON, Princeton University
The Underclass Concept: Self-Help vs. Government Intervention
WAYNE VROMAN, The Urban Institute
Transfer Payments, Sample Selection, and Male Black-White Earnings Differences
JEROME CULP, Duke University
Federal Courts and the Enforcement of Title VII
JONATHAN LEONARD, University of California-Berkeley
What was Affirmative Action?
Discussants: BERNARD ANDERSON, Princeton University
LYNN C. BURBRIDGE, The Urban Institute

10:15 A.M. *R&D*, INNOVATION, AND PUBLIC POLICY
Presiding: HENRY G. GRABOWSKI, Duke University
Papers: RICHARD NELSON, Yale University
Institutions Supporting *R&D* and the Role of Public Policy
EDWIN MANSFIELD, University of Pennsylvania
The *R&D* Tax Credit and Other Current Policy Issues
HENRY GRABOWSKI AND JOHN VERNON, Duke University
Regulation, Innovation and Imitation: Recent Legislative Initiatives
RICHARD LEVIN, Yale University
A New Look at the Role of the Patent System

Discussants: F. M. SCHERER, Swarthmore College
ROLF PIEKARZ, National Science Foundation

10:15 A.M. THE INTERNATIONAL DIMENSIONS OF FISCAL POLICIES
Presiding: JAMES TOBIN, Yale University
Papers: JACOB A. FRENKEL, University of Chicago and National Bureau of Economic Research, AND ASSAF RAZIN, Tel-Aviv University, Israel, and National Bureau of Economic Research
The International Transmission of Budget Deficits in the World Economy
JEFFREY SACHS, Harvard University and National Bureau of Economic Research
Fiscal Policies and Exchange Rates in the OECD
MARTIN S. FELDSTEIN, Harvard University and National Bureau of Economic Research
U.S. Fiscal Policies and the World Economy
Discussants: ROBERT J. GORDON, Northwestern University
WILLIAM H. BRANSON, Princeton University

10:15 A.M. SYNTHESIS THROUGH MICROANALYTIC SIMULATION
Presiding: GUY H. ORCUTT, Yale University
Papers: HEINZ P. GALLER, University of Bielefeld, FRG
The Frankfurt Microsimulation Model: A Comprehensive Approach to the Personal Income Distribution
GUNNAR ELIASSON, Industrial Institute of Economic and Social Research, Sweden
Pricing and Markets in the Swedish Microsimulation Model
ALICE NAKAMURA AND MASAO NAKAMURA, University of Alberta
Incorporation of Micro and Macro Information in Microanalytic Models
RICHARD RUGGLES AND NANCY RUGGLES, New Haven, Connecticut
Household Accounts and the Household Sector
Discussants: HAROLD BEEBOUT, Mathematica Policy Research, Inc.
STEVEN B. CALDWELL, Cornell University

10:15 A.M. SOME ISSUES IN INTERNATIONAL BANKING
Presiding: HEYWOOD FLEISIG, World Bank
Papers: GRACIELA KAMINSKY, University of California-San Diego
Capital Controls and the Dual Exchange Rate
MICHAEL H. MOFFETT, University of Oregon-Corvallis
The Stability of the International Interbank Market Under Credit Rationing
JOHN DUKAS AND ABOLHASSAN JALILVAND, Concordia University
Sovereign Risk and International Lending Interactions
Discussants: NANCY MARION, Dartmouth College
LANCE GIRTON, University of Utah
FRANK MCCORMICK, Bank of America

10:15 A.M. INTERGOVERNMENTAL FISCAL RELATIONS
Presiding: SHARON BERNSTEIN-MEGDAL, University of Arizona
Papers: THOMAS POGUE, University of Iowa
Intergovernmental Grants and Local Government Spending: Some New Evidence
DANIEL P. SCHWALLIE, Case Western Reserve University
A Theory of Intergovernmental Grants and their Effect on Aggregate Grantor-Recipient Spending
HOWARD CHERNICK, Hunter College, AND ANDREW RESCHOVSKY, Tufts University
State-Local Finance and the Flat Tax
Discussants: ROBERT PAUL INMAN, University of Pennsylvania
HELEN F. LADD, Harvard University

10:15 A.M. INVENTORIES AND EXPECTATIONS (Joint Session with the International Society for Inventory Research)
Presiding: MICHAEL C. LOVELL, Wesleyan University
Papers: F. OWEN IRVINE, JR., Michigan State University
Manufacturing Inventory Investment in the United Kingdom
MOHEB A. GHALI, University of Hawaii-Manoa
Seasonality, Aggregation and the Testing of the Production Smoothing Hypothesis
JOHN F. MUTH, Indiana University
Short-Run Forecasts of Business Activity
Discussants: RICCARDO FIORITO, University of Chieti, Italy
LOUIS J. MACCINI, The Johns Hopkins University
TRYPHON KOLLINTZAS, University of Pittsburgh

10:15 A.M. THE POLITICAL ECONOMY OF GROWTH, DISTRIBUTION, AND EQUITY
Presiding: ROBERT HEILBRONER, New School for Social Research
Papers: DONALD HARRIS, Stanford University
Growth and Equity: Complements or Opposites?
RICHARD MCGAHEY, New York University
The Political Economy of U.S. Growth and Distribution: 1960–84 and Beyond
LESTER THUROW, Masschusetts Institute of Technology
International Trade and the American Distribution of Income
Discussants: ROBERT KUTTNER, *The New Republic*
GLENN LOURY, Harvard University
ROBERT HEILBRONER, New School for Social Research

10:15 A.M. OMICRON DELTA EPSILON CHAPTER ADVISORS AND REGIONAL DIRECTORS (Joint Session with Omicron Delta Epsilon)
Presiding: DAVID A. MARTIN, State University of New York-Geneseo
(Papers and titles to be announced)

10:15 A.M. IMPACTS OF NEW MEDICAL CARE PAYMENT METHODS (Joint Session with the Health Economics Research Organization)
Presiding: ALLEN DOBSON, Health Care Financing Administration
Papers: BARRY KEATING, University of Notre Dame, AND HOWARD ADDIS, Gerig and Addis, Surgeons, Inc.
Prospective Payments: Will DRG Compensation Reduce Medical Costs?
ROBERT H. LEE, University of North Carolina
Prospective Reimbursement and the Demand for Medical Residents
DAVID I. KASS, U.S. General Accounting Office
Alternative Payment Mechanisms for Home Health Agencies
Discussants: JUDITH LAVE, University of Pittsburgh
MICHAEL A. MORRISEY, American Hospital Association
ZACHARY DYCKMAN, Center for Health Policy Studies

10:15 A.M. DEVELOPING COUNTRY ISSUES: THE ENERGY PROBLEM (Joint Session with the East-West Center)
Presiding: CORAZON M. SIDDAYAO, East-West Center
Papers: MARCIA GOWEN, East-West Center and World Bank
A Comparative Economic Analysis of Biomass and Fossil Fuels in Developing Countries
MOHAMMED KHALED, JULIE RANADA, K. RAVI SHANKAR, SUPAVUD SAICHEUA, AND CORAZON M. SIDDAYAO, East-West Center and University of Hawaii
A Comparison of Energy and Non-Energy Input Elasticities in the Manufacturing Sectors of Selected Asian Countries: Policy Implications
YUKO YANO, United Nations
Application of Microcomputer Technology for Energy Planning in Developing Countries
Discussants: PATRICIA KOSHEL, U.S. Agency for International Development
ERNST R. BERNDT, Massachusetts Institute of Technology
JOY DUNKERLEY, Resources for the Future

12:30 P.M. AEA CENTENNIAL LUNCHEON: THE FUTURE OF ECONOMICS
Presiding: ALICE M. RIVLIN, The Brookings Institution
Speakers: JOSEPH E. STIGLITZ, Princeton University
An Academy Viewpoint
MARINA V.N. WHITMAN, General Motors
A Business Viewpoint
RUDOLPH G. PENNER, Congressional Budget Office
A Government Policy Viewpoint

2:30 P.M. SUPPLY-SIDE ECONOMICS: WHAT REMAINS?
Presiding: HERBERT STEIN, American Enterprise Institute
Papers: MARTIN FELDSTEIN, Harvard University and National Bureau of Economic Research
Supply-Side Considerations and the Policy Mix
LAWRENCE CHIMERINE, Chase Econometrics
Economic Surprises and Messages of the 1980's
GEORGE M. VON FURSTENBERG AND R. JEFFERY GREEN, Indiana University
Supply-Side Modeling from Bits and Pieces
Discussants: BARRY P. BOSWORTH, The Brookings Institution
MANUEL H. JOHNSON, U.S. Department of the Treasury
VICTOR A. CANTO, University of Southern California

2:30 P.M. EQUITY BETWEEN THE SEXES IN ECONOMIC PARTICIPATION
Presiding: IRENE LURIE, State University of New York-Albany
Papers: SOPHIE KORCZYK, Employee Benefit Research Institute, Washington, D.C.
Tax Reform and Equity for the Two-Earner Family: The Problem of Employee Benefits
ELAINE SORENSEN, University of Massachusetts
Implementing Comparable Worth: A Survey of Recent Job Evaluation Studies
MICHELLE WHITE, University of Michigan
Commuting Behavior of Men and Women Workers in Large Cities
NAN L. MAXWELL, California State University-Hayward AND RONALD D'AMICO, Ohio State University
Employment and Wage Effects of Involuntary Job Separation: Male-Female Differences

2:30 P.M. THE SOVIET GROWTH SLOWDOWN: THREE VIEWS
Presiding: HERBERT S. LEVINE, University of Pennsylvania
Papers: STANISLOW GOMULKA, London School of Economics
Soviet Growth Slowdown: Duality, Maturity, and Innovation
PADMA DESAI, Columbia University
Soviet Growth Slowdown and Efficiency
VLADIMIR KONTOROVICH, Command Economies Research, Inc.
Soviet Growth Slowdown: Econometric vs. Direct Evidence
Discussants: EDWARD F. DENISON, The Brookings Institution
GERTRUDE SCHROEDER, University of Virginia

2:30 P.M. SITING OF HAZARDOUS FACILITIES
Presiding: HOWARD KUNREUTHER, University of Pennsylvania
Papers: ROBERT CAMERON MITCHELL, Resources for the Future, AND RICHARD T. CARSON, Resources for the Future and University of California-San Diego
Not in My Backyard: The Economic Cost of Siting Opposition
V. KERRY SMITH, Vanderbilt University, AND WILLIAM H. DESVOUSGES, Research Triangle Institute
Asymmetries in the Valuation of Risk and the Siting of Hazardous Waste Facilities
HOWARD KUNREUTHER AND PAUL R. KLEINDORFER, University of Pennsylvania
The Rate of Compensation and Insurance in Siting Hazardous Facilities
Discussants: ALLEN V. KNEESE, Resources for the Future
WILLIAM D. SCHULZE, University of Colorado

2:30 P.M. BUDGET REFORM AND THE THEORY OF FISCAL FEDERALISM
Presiding: JOHN M. QUIGLEY, University of California-Berkeley
Papers: MANCUR OLSON, University of Maryland
The Continuing Evolution of Fiscal Federalism
HARVEY S. ROSEN, Princeton University
State Taxation in a Federal System
JOHN M. QUIGLEY AND DANIEL L. RUBINFELD, University of California-Berkeley
Budget Reform and the Structure of American Federalism
Discussants: RUDOLPH PENNER, U.S. Congressional Budget Office
RICHARD MUSGRAVE, University of California-Santa Cruz
GORDON TULLOCK, George Mason University

2:30 P.M. THE SURVEY OF INCOME AND PROGRAM PARTICIPATION—EARLY FINDINGS FROM A NEW DATA RESOURCE
Presiding: MARTIN DAVID, University of Wisconsin-Madison
Papers: PAUL RYSCAVAGE, U.S. Bureau of the Census
Labor Force Transitions
ENRIQUE J. LAMAS AND JOHN M. MCNEIL, U.S. Bureau of the Census
Patterns of Asset Ownership and Wealth Holdings
CONSTANCE F. CITRO, Mathematica Policy Research, AND HAROLD W. WATTS, Columbia University
Patterns of Household Composition and Family Status Change
Discussants: GLEN C. CAIN, University of Wisconsin
JAMES TOBIN, Yale University

2:30 P.M. ISSUES IN FINANCIAL REFORM
Presiding: GILLIAN GARCIA, U.S. General Accounting Office
Papers: IAN GIDDY, Columbia University and New York University
Risk and the Underwriting of Corporate Securities
HERBERT BAER AND LARRY MOTE, Federal Reserve Bank of Chicago
Lessons from Foreign Experience with Nationwide Banking

ELIJAH BREWER, Federal Reserve Bank of Chicago, GILLIAN GARCIA, U.S. General Accounting Office, AND ALAN REICHERT, Northern Illinois University
The Determinants of S&L Profitability: An Intertemporal Comparison
Discussants: SUSAN KRAUSE, Office of the Comptroller of the Currency
CONSTANCE DURHAM, Federal Reserve Bank of Boston
GEORGE KAUFMAN, Loyola University of Chicago
JOHN F. SMITH, Loyola University of Chicago

2:30 P.M. IS BETTER FORECASTING POSSIBLE?
Presiding: ALBERT T. SOMMERS, Conference Board
Papers: VICTOR ZARNOWITZ, University of Chicago
The Forecasting Record
PAUL MCCRACKEN, University of Michigan
Views of a Policymaker
LAWRENCE CHIMERINE, Chase Econometrics
Views of a Practitioner

2:30 P.M. MARKET STRUCTURE AND COMPETITION
Presiding: JAMES W. MCKIE, University of Texas-Austin
Papers: FREDERICK R. WARREN-BOULTON, ROBERT H. MCGUCKIN, AND PETER WALSTEIN, U.S. Department of Justice
Analysis of the Competitive Effects of Mergers with an Extension to the Capital Asset Pricing Model
STANLEY S. REYNOLDS, University of Arizona
Plant Closings and Exit Behavior in Declining Industries
JOHN KWOKA, George Washington University
Market Segmentation by Product Quality: Some Evidence from Automobiles
Discussants: LEE E. PRESTON, University of Maryland
STEVEN N. WIGGINS, Texas A&M University

2:30 P.M. INSTITUTIONS AND POLICIES IN HIGHER EDUCATION
Presiding: DAVID BRENEMAN, Kalamazoo College
Papers: STEPHEN A. HOENACK, University of Minnesota
An Econometric Model of a Public University's Input and Output
ROBERT C. DOLAN AND ROBERT M. SCHMIDT, University of Richmond
Modeling Institutional Production of Higher Education
SANDRA R. BAUM, Wellesley College, AND SAUL SCHWARTZ, Tufts University
Merit-Based vs. Need-Based Aid for Post-Secondary Students
MICHAEL S. MCPHERSON, The Brookings Institution
Institutional Behavior and Student-Aid Policy
Discussants: (To be announced)

2:30 P.M. ECONOMIC IMPLICATIONS OF FAMILY AND TRADITIONAL TIES IN DEVELOPING COUNTRY LABOR MARKETS (Joint Session with the Industrial Relations Research Association)
Presiding: BERNARD E. ANDERSON, Princeton University
Papers: SUBBIAH KANNAPPAN, Michigan State University
Maximization and Labor Market Behavior in an Anthropological and Political Context
KOJI TAIRA, University of Illinois
Diverse Entrepreneurial Traditions and Implications for Internal and External Labor Markets
JAMES G. SCOVILLE, University of Minnesota
Economic Analysis of Labor Market Premia for Traditional Skills in Developing Nations
Discussants: T. PAUL SCHULTZ, Yale University
MICHAEL PIORE, Massachusetts Institute of Technology
MARK LEISERSON, World Bank

2:30 P.M. ECONOMIC RESPONSES TO PENSION PROVISIONS
Presiding: ALICIA H. MUNNELL, Federal Reserve Bank of Boston
Papers: ROBERT B. AVERY, GREG ELLIEHAUSEN, Board of Governors of the Federal Reserve System, AND THOMAS GUSTAFSON, Health Care Financing Administration
Household Responses to Pension Coverage from the 1983 *Survey of the Consumer Finances*
RICHARD V. BURKHAUSER, Vanderbilt University, AND KAREN C. HOLDEN, University of Wisconsin
Pensioners' Widows: Are They Shortchanged?
R. GLENN HUBBARD AND KENNETH L. JUDD, Northwestern University
Social Security and Individual Welfare: Precautionary Saving, Borrowing Constraints, and Optimal Taxation

ROBERT L. CLARK AND ANN A. MCDERMED, North Carolina State University
Earnings Response to Pension Coverage and Eligibility
Discussants: OLIVIA S. MITCHELL, Cornell University
RICHARD A. IPPOLITO, U.S. Department of Labor

2:30 P.M. PROVISION OF LONG-TERM MEDICAL CARE (Joint Session with the Health Economics Research Organization)
Presiding: WILLIAM J. SCANLON, Georgetown University
Papers: STEVEN G. ULLMANN AND A. G. HOLTMANN, University of Miami
Insight into the Demand for and Utilization of Long-Term Health Care Service in the United States
NICCIE L. MCKAY, Texas A&M University
Costs and Input Choices in the Nursing Home Industry
PETER KEMPER, RANDALL BROWN, JON CHRISTIANSON, THOMAS GRANNEMANN, AND JUDITH WOOLRIDGE, Mathematica Policy Research, Inc.
Results of the National Long-Term Care Channeling Demonstration
Discussants: ROBERT SCHLENKER, University of Colorado
LYNN PARRINGER, California State University-Hayward
WILLIAM POLLAK, University of Chicago

4:30 P.M. PRESIDENTIAL ADDRESS AND BUSINESS MEETING
Presiding: ALICE M. RIVLIN, The Brookings Institution
Speaker: CHARLES P. KINDLEBERGER, Massachusetts Institute of Technology
International Public Goods without International Government

6:30 P.M. AEA 100TH BIRTHDAY CELEBRATION

Monday, December 30, 1985

8:00 A.M. GRANTS ECONOMY AND COMPARATIVE ECONOMIC SYSTEMS: MEASUREMENT PROBLEMS AND POLICY ISSUES (Joint Session with the Association for the Study of the Grants Economy)
Presiding: JANOS HORVATH, Butler University
Papers: ALAN A. BROWN, University of Windsor, AND EGON NEUBERGER, State University of New York-Stony Brook
On the Efficiency and Normative Effects of Grant Elements in the Market Sphere
ELIZABETH CLAYTON, University of Illinois and University of Missouri
The Impact of Subsidies on the Supply Function in Soviet Agriculture
Discussants: DEBORAH D. MILANKOVITCH, Barnard College
LYNN TURGEON, Hofstra University

8:00 A.M. HIGH TECH INDUSTRIES AND ECONOMIC GROWTH (A ROUNDTABLE) (Joint Session with the National Association of Business Economists)
Presiding: EDGAR R. FIEDLER, The Conference Board
Panel: ROSANNE COLE, IBM Corporation
GEORGE M. VON FURSTENBERG, Indiana University
ULRIC WEIL, Morgan Stanley and Company

8:00 A.M. GOVERNMENT POLICY AND POVERTY
Presiding: PETER GOTTSCHALK, Bowdoin College
Papers: GLENN LOURY, Harvard University
Welfare Attitudes and Poverty
RICHARD MICHEL, The Urban Institute, AND FRANK LEVY, University of Maryland
Welfare Work and Poverty
SHELDON DANZIGER, University of Wisconsin-Madison, AND PETER GOTTSCHALK, Bowdoin College
Growth Cycles and Poverty
Discussants: GARY BURTLESS, The Brookings Institution
DAVID ELLWOOD, Harvard University

8:00 A.M. ECONOMIC POLICY AND THE THEORY OF THE FIRM: NEW PERSPECTIVES
Presiding: MICHAEL C. JENSEN, University of Rochester and Harvard University
Papers: SHERWIN ROSEN, University of Chicago
The Public Interest in Labor Contracts
STEVEN C. SALOP, Georgetown University
Limited Cooperation, Vertical Integration, and Competitive Incentives

OLIVER E. WILLIAMSON, Yale University
Transforming Merger Policy: The Pound of New Perspectives
Discussants: RONALD H. COASE, University of Chicago
GREGORY K. DOWN, Yale University
DAVID SAPPINGTON, Bell Communications Research and University of Pennsylvania

8:00 A.M. THE ECONOMICS OF PUBLIC INFORMATION, GOVERNMENT POLICY, AND MARKET EXPECTATIONS
Presiding: SUCHADA V. LANGLEY, University of Maryland
Papers: BARRY FALK AND PETER ORAZEM, Iowa State University
The Theory of Futures Market Responses to Government Crop Forecasts
DAVID BESSLER, Texas A&M University, CARL NELSON, University of Illinois, AND EDNA LOEHMAN, Purdue University
The Effect of Information on Input Use Decisions
WALLACE HUFFMAN, Iowa State University, AND SUCHADA LANGLEY, University of Maryland
The Differential Effects of Relative Expected Prices on the Farm and Non-Farm Sectors
Discussants: JOHN MIRANOWSKI, U.S. Department of Agriculture
ROBERT CHAMBERS, University of Maryland
ROBERT EVENSON, Yale University

8:00 A.M. INTERNATIONAL TRADE AND THE *LDC*s
Presiding: PAUL WONNACOTT, University of Maryland
Papers: WING T. WOO, University of California-Davis
The Impact of the U.S. Policy Mix on the *LDC*s
MONTAGUE J. LORD, Inter-American Development Bank
Product Differentiation in International Commodity Trade
BETH V. YARBROUGH, Amherst College, AND ROBERT M. YARBROUGH, Tufts University
Reciprocity, Bilateralism, and Regulatory Standards: Protectionism or Hostages to Free Trade?
Discussants: STEPHEN A. O'CONNELL, University of Pennsylvania
ERIC FISHER, Board of Governors of the Federal Reserve System
MIGUEL A. KIGUEL, University of Maryland

8:00 A.M. MACROECONOMIC ADJUSTMENT PROCESSES
Presiding: VAN DOORN OOMS, Committee on the Budget, U.S. House of Representatives
Papers: JOHN CASKEY AND STEVEN FAZZARI, Washington University-St. Louis
Disinflation, Wage Flexibility, and Nominal Debt Commitments
IAN DOMOWITZ, R. GLENN HUBBARD, AND BRUCE C. PETERSEN, Northwestern University
Industry Margins and the Business Cycle
MARK WILSON, University of Pennsylvania
The Response of Manufacturing Labor to De-Industrialization and the Growth of Service Employment
Discussants: FRANK DE LEEUW, U.S. Department of Commerce
MICHAEL C. LOVELL, Wesleyan University

8:00 A.M. RACE AND ETHNICITY IN THE HISTORY OF ECONOMIC THOUGHT (Joint Session with the National Economic Association)
Presiding: BOBBIE L. HORN, University of Tulsa
Papers: MARK ALDRICH, Smith College
Race, Ethnicity and the Thought of John R. Commons
ROBERT CHERRY, Brooklyn College
Walker to Myrdal: The Transition from Genetic to Culture of Poverty Explanations of Black Inferiority
STEVEN SHULMAN, Colorado State University
Alternative Conceptualizations of Labor Market Discrimination
Discussants: FRANCIEL WILSON, University of Michigan
HOWARD STANBACK, New School for Social Research

8:00 A.M. POPPER AND THE LSE ECONOMISTS (Joint Session with the History of Economics Society)
Presiding: E. ROY WEINTRAUB, Duke University
Papers: NEIL DE MARCHI, Duke University
Popper and the LSE
D. WADE HANDS, University of Puget Sound
Ad hocness in Economics
DAN HAUSMAN, Carngie-Mellon University
Popper Misapprehended

DALE STAHL, Duke University and Massachusetts Institute of Technology
Testing and the Information Content of Theory
Discussants: BRUCE CALDWELL, University of North Carolina-Greensboro
ALEXANDER ROSENBERG, Syracuse University
CHRIS ARCHIBALD, University of British Columbia
LAWRENCE A. BOLAND, Simon Fraser University
BRUCE CALDWELL, University of North Carolina-Greensboro

8:00 A.M. FUTURE PROSPECTS OF THE INDIAN ECONOMY (Joint Session with the Association of Indian Economic Studies)
Presiding: ROMESH DIWAN, Rensselaer Polytechnic Institute
Papers: J. D. SETHI, Delhi University, India
Future Prospects of the Indian Economy
F. TOMASSON JANNUZI, University of Texas-Austin
Toward Food Security in India
ROMESH DIWAN, Rensselaer Polytechnic Institute
Will the Indian Economy Take Off?
Discussants: SURESH DESAI, Montclair State College
ASHOK BHARGAVA, University of Wisconsin-Whitewater

8:00 A.M. WORLD POPULATION TRENDS AND THEIR IMPACT ON DEVELOPMENT (Joint Session with the Society for Policy Modeling)
Presiding: DOMINICK SALVATORE, Fordham University
Papers: WUU-LONG LIN, United Nations
Socio-Demographic Characteristics of Economic Disparities
FRED COMPANO, United Nations
The Process of Urbanization and its Effects on Income Distribution
GEORGE STOLNITZ, Indiana University
Population and Development of Interrelationships: Main Food and Agricultural Issues
Discussants: DOUGLAS O. WALKER, United Nations
STANLEY J. LAWSON, St. John's University
GEORGE C. WANG, California State University

8:00 A.M. THE DEMAND FOR PHYSICIAN SERVICES (JOINT SESSION WITH THE HEALTH ECONOMICS RESEARCH ORGANIZATION)
Presiding: WILLIAM D. MARDER, American Medical Association
Papers: BONG-MIN YANG, University of Lowell, AND TEH-WEI HU, Pennsylvania State University
Demand and Supply of Physician Services: Equilibrium vs. Disequilibrium Analysis
JOACHIM NEIPP, Harvard University and Heidelberg University
Supplier Induced Demand: A Refined Theory and Test
ALAN C. MONHEIT AND CURT D. MUELLER, National Center for Health Services Research
Demand for Primary Care Physicians' Services: An Hedonic Approach
Discussants: PETER COYTE, University of Alberta
JAMES B. RAMSEY, New York University
MICHAEL GROSSMAN, City University of New York

10:15 A.M. INFLATION, UNEMPLOYMENT, AND FIRM FINANCE IN HISTORICAL PERSPECTIVE
Presiding: CLAUDIA GOLDIN, University of Pennsylvania
Papers: STEVEN B. WEBB, University of Michigan
Four Ends of the Big Inflation in Germany, 1920 to 1924
JOHN A. JAMES, University of Virginia
Shifts in the Nineteenth-Century Phillips Curve
RAYMOND P. H. FISHE AND LOUIS DE ALESSI, University of Miami
Why Do Companies Distribute Assets?
Discussants: HUGH ROCKOFF, Rutgers University
MICHAEL EDELSTEIN, Queens College

10:15 A.M. ISSUES IN ECONOMIC DEVELOPMENT
Presiding: LARRY E. WESTPHAL, Swarthmore College
Papers: RICARDO MARTIN AND MARCELO SELOWSKY, World Bank
External Shocks and the Demand for Adjustment Finance
KRISHNA BELBASE AND RICHARD GRABOWSKI, Southern Illinois University-Carbondale
An Analysis of Optimal Scale and Factor Intensity in Nepalese Agriculture: An Application of Ray-Homothetic Production Function

JENE KWON, Northern Illinois University
An Estimation of Production Technologies in South Korean Manufacturing
Discussants: CLIVE BELL, The Johns Hopkins University
DAVID BLAU, University of North Carolina-Chapel Hill

10:15 A.M. THE EFFECTS OF TRANSPORTATION DEREGULATION: THEORY AND EVIDENCE
Presiding: THEODORE E. KEELER, University of California-Berkeley
Papers: JOEL DERNSKI, Stanford University, DAVID SAPPINGTON, Bell Communications Research, AND PABLO SPILLER, The Hoover Institution
Regulation and Competition: A Model of Endogenous Regulatory Change
RICHARD BEILOCK AND JAMES FROEMAN, University of Florida
The Impact on Freight Rates of Motor Carrier Deregulation: The Cases of Florida and Arizona
FREDERICK HARRIS AND JOHN TRAPANI, University of Texas-Arlington
Firm-Specific Price and Quality in Competitive Airline Markets
Discussants: DARIUS W. GASKINS, JR., Burlington Northern Railroad
MICHAEL E. LEVINE, University of Southern California

10:15 A.M. ASIAN ECONOMIC STUDIES: KOREA (Joint Session with the Committee on Asian Economic Studies)
Presiding: LAWRENCE B. KRAUSE, The Brookings Institution
Papers: SUNG Y. KWACK, Howard University
Dynamic Effects of a Change in Savings Rates on Growth and External Debt: The Case of Korea
KISEOK LEE AND MANORANJAN DUTTA, Rutgers University
The Structural Instability of International Trade—Korea, Japan, the United States, and the Rest of the World
JANG H. YOO, Virginia Commonwealth University, AND BONG J. YOON, State University of New York-Binghamton
The Demand for Urban Housing in Korea: Rent Seeking and Welfare Transfer
Discussants: RICHARD HOOLEY, University of Pittsburgh
DAVID COLE, Harvard Institute for International Development
SHUJIRO URATA, World Bank

10:15 A.M. THE MYSTIFYING ENTREPRENEURSHIP: ON THE RELATION BETWEEN ENTREPRENEURSHIP, MOTIVATION, AND E-EFFICIENCY THEORY (Joint Session with the Society for the Advancement of Behavioral Economics)
Presiding: BENJAMIN GILAD, Rutgers University
Papers: HARVEY LEIBENSTEIN, Harvard University
Entrepreneurship, Motivation, and X-Efficiency Theory
PHILIP B. NELSON AND JOHN HILKE, Federal Trade Commission
Caveat Innovator: Strategic and Structural Implications of Consumer Shopping Behavior for New Product Introduction
WILLIAM STARBUCK, New York University
Whatever Happened to the Behavioral Theory of the Firm?
(Winner of the SABE Competition)
(Title to be announced)
Discussants: STANLEY KAISH, Rutgers University
SHLOMO MAITAL, Technion, Israel

10:15 A.M. CONFLICT AND PEACE ECONOMICS II: NATIONAL SECURITY AND MILITARY EXPENDITURES (Joint Session with the Peace Science Society (International))
Presiding: KENNETH BOULDING, George Mason University and University of Colorado-Boulder
Papers: RICHARD NELSON, Yale University
Issues in Governance of National Security
CHARLES ANDERTON AND WALTER ISARD, Cornell University
Arms Control: Economic Constraints and Stimuli
Discussants: MICHAEL INTRILIGATOR, University of California-Los Angeles
MURRAY WOLFSON, Oregon State University

10:15 A.M. STRUCTURAL TRENDS IN TRANSPORTATION AND PUBLIC UTILITIES (Joint Session with the Transportation and Public Utilities Group)
Presiding: JOHN C. SPYCHALSKI, Pennsylvania State University
Papers: HARRY M. TREBING, Michigan State University
Emergent Patterns of Market Dominance in Energy and Communications: Fallacies in the New Anti-Regulatory Economics

ALICE E. KIDDER, Babson College
Carrier Turnover in Rural Transportation Markets
ROBERT N. STEARNS, Barber Associates
Mergers and Acquisitions in the Rail Industry
Discussants: KENNETH R. NOWOTNY, New Mexico State University
ALLAN D. SCHUSTER, University of Texas-Austin

10:15 A.M. OUTLOOK FOR THE FINANCIAL MARKETS (Joint Session with the National Association of Business Economists)
Presiding: ANNE D. PICKER, Nynex Corporation
Papers: DANIEL AHEARN, Wellington Management
Interest Rates
GEOFFREY BELL, Geoffrey Bell and Company, Inc.
Foreign Exchange
PETER VERMILYE, Endowment Management and Research Corporation
Security Markets
Discussants: GEORGE MCKINNEY, JR., University of Virginia
FRANK SCHOTT, Equitable Life Assurance Company

10:15 A.M. THE DETERMINANTS OF NATURAL RESOURCE SCARCITY (Joint Session with the Association of Environmental and Resource Economists)
Presiding: RICHARD NORGAARD, University of California-Berkeley
Papers: ERNST R. BERNDT, Massachusetts Institute of Technology
The Role of Substitution and Conservation in the Demand for Energy
ROBERT HALVORSEN, University of Washington
Evidence on Resource Scarcity from Minerals Markets
ROGER SEDJO, Resources for the Future, Inc.
Deforestation and the Value of Biological Diversity
Discussants: EDWARD MOREY, University of Colorado
SCOTT FARROW, Carnegie-Mellon University
DOUGLAS SOUTHGATE, Ohio State University

10:15 A.M. ANALYSES OF THE NATION'S LARGEST HEALTH CARE SYSTEM (Joint Session with the Health Economics Research Organization)
Presiding: JERRY CROMWELL, Center for Health Economics Research
Papers: BERNARD FRIEDMAN, ANDREW MELCZER, Northwestern University, AND ROBERTA L. DENWOOD, Edward Hines, Jr. Hospital
Resources Used in Ambulatory Care in VA Hospitals
EDGAR A. PEDEN, Veterans' Administration
Determinants of VA and Other Hospital Use
DONALD E. YETT AND RICHARD L. ERNST, University of Southern California
VA Medical Care Demand Models
Discussants: CRAIG COELAN, Abt Associates, Inc.
AMY TAYLOR, National Center for Health Services Research
JACK HADLEY, Georgetown University

10:15 A.M. ECONOMIC ISSUES IN THE ARTS
Presiding: MANCUR OLSON, University of Maryland
Papers: WILLIAM J. BAUMOL, Princeton University and New York University
Investment in the Visual Arts as a Floating Crap Game
DICK NETZER, New York University
Dance in New York: Market and Subsidy Changes
JAMES H. GAPINSKI, Florida State University
The Lively Arts as Substitutes for the Lively Arts
Discussants: WILLIAM S. HENDON, *Journal of Cultural Economics*
HAROLD HOROWITZ, National Endowment for the Arts
VIRGINIA LEE OWEN, Illinois State University

NOTES

The ninety-eighth annual meeting of the American Economic Association will be held in New York, New York, December 28–30, 1985.

The Professional Placement Service will be located at the Sheraton Centre Hotel. It will be open from 10:00 A.M. to 5:00 P.M., December 27; 9:00 A.M. to 5:00 P.M., December 28–29, and 9:00 to 12:00 noon, December 30.

Members wishing to give papers or make suggestions for the program for the AEA meeting, New Orleans, LA, December 28–30, 1986, are invited to write to Professor Gary S. Becker, Department of Economics, 1126 East 59th Street, University of Chicago, Chicago, IL 60637. To be considered for contributed sessions, abstracts of (noneconometric) papers must be received no later than February 1, 1986.

The Lionel Robbins Memorial Fund: After the death of Lionel Robbins last year, friends, colleagues, and former students established a fund to honor his memory. The fund will be used for two purposes. The first is sponsorship of a set of lectures by outstanding economists to be presented annually at the London School of Economics. Second, it will provide post-graduate research scholarships open to students of all nationalities and usable at any U.K. institution of higher learning. It is hoped to collect the equivalent of at least $250,000 overall, and about half the amount has already been contributed in the United Kingdom. Contributions toward the lecture program can be made through the American Friends of L.S.E., and are tax deductible. Checks should be made payable to them, marked "Lionel Robbins Lecture Series," and sent to American Friends of L.S.E., 1497 Chain Bridge Road, Suite 104, McLena, VA 22101. Contributions toward the scholarships, which may not be tax deductible, should be sent directly to the Appeals Officer, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, England.

The P.W.S. Andrews Memorial Prize of £300 is awarded annually for an essay within the general field of Industrial Economics and Theory of the Firm, broadly interpreted. Entry is open to those under age 30, or within eight years of taking a first degree. The essay, written in English and not exceeding 10,000 words, should be original research by the candidate only, not previously published, and not previously awarded any other prize. The winning essay may, at the discretion of the editors, be published in *The Journal of Industrial Economics*. The deadline is December 31, 1985. Further details are available from Miss B. Cox, University House, The University of Lancaster, Bailrigg, Lancaster, LA1 4YW, England, or the General Editor, *JIE*, The Institute of Economics and Statistics, The University of Oxford, St. Cross Bldg., Manor Road, Oxford, OX1 3UL, England, or the North American Editor, *JIE*, Graduate School of Business Administration, New York University, 90 Trinity Place, New York, NY 10006.

Nominations for candidates for Sloan Research Fellowships are due by September 15. Candidates must be members of the faculty at a college or university in the United States or Canada, and must be at an early stage of their research careers. Please write Sloan Research Fellowships, Alfred P. Sloan Foundation, 630 Fifth Avenue, New York, NY 10111.

The Leonard J. Savage Award of $500 is presented annually for an outstanding doctoral dissertation in the area of Bayesian Econometrics and Statistics. The closing date each year is July 1. Dissertations and supporting letters from thesis supervisors should be sent to Professor Arnold Zellner, Graduate School of Business, University of Chicago, 1101 E. 58th Street, Chicago, IL 60637.

The co-winners of the 1984 Award are Luc Bauwens, "Bayesian Full Information Analysis of Simultaneous Equation Models Using Integration by Monte Carlo," completed at CORE, Université Catholique de Louvain, Louvain, Belgium, and Peter C. Cramton, "The Role of Time and Information in Bargaining," completed at the Graduate School of Business, Stanford University. The Thesis Evaluation Committee included Bruce M. Hill, University of Michigan, chair; James M. Dickey, State University of New York; Nicholas M. Kiefer, Cornell University; David Lane, University of Minnesota; and Arnold Zellner, University of Chicago (*ex officio*).

The Robert Wood Johnson Foundation announces a fellowship program for university faculty in the field of health care finance. The program has three parts: a four-month educational component at The Johns Hopkins University; an eighth-month placement in a private or public health care financing organization; and support for related research in the year following. For more information and application materials, contact the Program Director, Dr. Carl J. Schramm, The Center for Hospital Finance and Management, The Johns Hopkins Medical Institutions, 624 North Broadway, Baltimore, MD 21205 (telephone 301 + 955–8316).

The National Humanities Center will admit approximately forty Fellows for the academic year, 1986–87. Fellowships are awarded on the basis of open competition, and applications are welcomed from the United States and abroad. Representatives of the humanities, natural sciences, social sciences, and professional life may apply. The deadline is October 15, 1985. Application forms are available from the Center and should be supported by a curriculum vitae, a 1,000-word project proposal, and three letters of recommendation. Please write Kent Mullikin, Assistant Director, National Humanities Center, 7 Alexander Drive, Research Triangle Park, NC 27709.

The Institute for Research on Poverty at the University of Wisconsin, in association with the Office of the Assistant Secretary for Planning and Evaluation at the U.S. Department of Health and Human Services, is sponsoring a Small Grants Program for research on a variety of poverty-related topics. Program guidelines will be available November 1, 1985. The grants will fund research for the summer of 1986; the application deadline is February 15, 1986. Further information on obtaining program guidelines can be obtained from Elizabeth Evanson, Institute for Research on Poverty, 1180 Observatory Drive, 3412 Social Science Building, University of Wisconsin, Madison, WI 53706.

Applications are invited for the Graduate School of Industrial Administration, Carnegie-Mellon University, postdoctoral fellowships in political economy. Recent Ph.D.s (within past four years) in economics, political science, or related fields with a strong commitment to the use of mathematical or quantitative analysis in the study of politics and the interdependence of political and economic decision making should apply. Special consideration will be given to scholars interested in pursuing collaborative research with GSIA faculty in studying political processes that relate to government spending and taxation, and in studying legislatures and voting. A twelve-month stipend commensurate with an assistant professor's salary will be provided. Please forward a resume and brief statement of research interests, or write for further information: Professor Howard Rosenthal, GSIA, Carnegie-Mellon University, Pittsburgh, PA 15213.

Donation Available: Full issues of the *American Economic Review* (1958–85) and *Journal of Economic Literature* (1966–85) for only shipping charges to any university or college. Contact A. J. Harrison, 27 West 55 Street, New York, NY 10019 (telephone 718+990–0623).

The NSF Division of Science Resources Studies announces its 1986 Program for the Analysis of Science Resources. Research grants will be awarded for studies directed toward quantitative indicators for science and technology policy within such areas as New Institutional Arrangements for Promoting *R&D* Technological Innovation and Diffusion; Bibliometric and other Measures of Scientific or Technological Advance; International Flows of Science and Technology, including Personnel; Public Attitudes toward Science and Technology; and the Relationships of Science and Technology with Society. Proposals should be submitted by January 15, 1986. Awards usually cover one to two person-years of effort. Specific information is given in the Announcement of the Program, which can be obtained from: Division of Science Resources Studies, NSF, 1800 G Street, NW, Washington, D.C. 20550.

New annual grant of up to $50,000 is announced by the law firm, O'Melveny & Meyers. The 1986 topic is the Internationalization of American Business—The Challenge of the Pacific Basin. Applications (by post-doctoral faculty, businessmen, other professionals) must be on file by October 1, 1985. The grantee for 1986 will be selected by a committee consisting of John F. Akers, President and Chief Executive Officer of IBM; Bruce MacLaury, President of The Brookings Institution; John McArthur, Dean of the Harvard Graduate School of Business; and Susan Westerberg Prager, Dean of UCLA Law School and President Elect of the Association of American Law Schools. Application forms may be obtained by writing to Warren Christopher, O'Melveny & Meyers, 400 South Hope Street, Los Angeles, CA 90071–2899.

Call for Papers: Studies in Economic Analysis is a biannual, student-edited journal soliciting research articles from both established economists and students. Submission fee is $4.00 for nonsubscribers. Submit manuscripts to, or request format and style instructions from, The Editors *SEA*, Department of Economics, College of Business Administration, University of South Carolina, Columbia, SC 29208.

Call for Papers: The Western Social Science Association will hold its 28th annual meeting at the MGM Grand Hotel, Reno, Nevada, April 23–26, 1986. Abstracts or proposals for mainstream economics should be sent to WSSA President Don Coursey, Department of Economics, University of Wyoming, Laramie, WY 82701 prior to November 15, 1985.

Call for Papers: VOICES: The Art and Science of Psychotherapy is the quarterly journal of the American Academy of Psychotherapists. A special issue to appear fall 1986 is being planned on Ploughing the Outer Fields: Anthropologists, Sociologists, and Economists; Newer Contributions to Psychotherapy. Papers both informal and informative are welcome. They should be approximately 2500–3000 words, double spaced, submitted in triplicate. The deadline for submissions is July

1, 1986. Please send papers to Dr. E. Mark Stern, Editor, *VOICES*, 215 East 11 Street, New York, NY 10003.

Call for Papers: The seventh Berkshire Conference on the History of Women will be held June 19–21, 1987, at Wellesley College. The theme is "Beyond the Public/Private Dichotomy: Reassessing Women's Place in History." Please submit proposals for complete sessions (no more than two papers, one commentator, and one moderator). Individual papers will also be considered. Please include title of paper(s), one-page abstract(s), names, addresses, telephone numbers, and one-page vitaes of all participants to include affiliation or place of residence. Send three copies (other than U.S. topics) to Dorothy O. Helly, Women's Studies Program, Hunter College, 695 Park Avenue, New York, NY 10021; (U.S. topics) to Susan Reverby, Women's Studies Program, Wellesley College, Wellesley, MA 02181. The deadline is February 1, 1986. Enclose a stamped self-addressed postcard.

Call for Papers: A Risk Theory Seminar will be held at the University of South Carolina, April 4–6, 1986. Topics on all areas of risk and insurance are eligible. Send outlines of proposed presentations by November 30, 1985 (eight or nine will be chosen) to any of the following: Dr. Barbara Beliveau, School of Business Administration, U-41B, University of Connecticut, Storrs, CT 06268; Dr. S. Travis Pritchett, College of Business Administration, University of South Carolina, Columbia, SC 29208; Dr. Harris Schlesinger, Department of Economics, Vanderbilt University, Nashville, TN 37235. The acceptance decision will be made by January 20, 1986; completed papers must be sent by February 20, 1986.

Call for Papers: The Southwestern Economics Association will hold its annual meeting at the Menger Hotel, San Antonio, Texas, March 19–22, 1986, in conjunction with the Southwest Social Science Association's annual meeting. All academic, business, and government economists are invited to submit proposals for papers and panel discussions, or to serve as chair or discussant. Send abstracts or other information before October 15, 1985, to M. Ray Perryman, Suite 212, Hankamer School of Business, Baylor University, Waco, TX 76798.

Call for Papers: The Southeastern Council of Latin American Studies 1986 meeting will be held at Clemson University, April 3–5. The theme is "City and Country in Latin American: The Implications for Change." Proposals for panels, papers, and commentators in all disciplines are invited. Contact the Program Chairpersons: Professor Bowdler, Political Science Department, University of South Carolina, Aiken, SC 29801 *and* Professor Kargleder, Department of Languages, Spring Hill College, Mobile, AL 36608. The deadline is October 15, 1985.

Call for Papers: The fourth International Conference on Cultural Economics and Planning, sponsored by the Ministry of Culture of France and the Association for Cultural Economics, will be held in Avignon, France, at the Palace of the Popes, May 5–6, 1986. Sessions will deal with economics of the arts, historic preservation, among other topics. To present a paper, or attend, contact Dr. William S. Hendon, Association for Cultural Economics, Department of Urban Studies, University of Akron, Akron, OH 44325.

Call for Papers: The 1986 annual meeting of the Eastern Finance Association will be held April 16–19, 1986 in Nashville, Tennessee. To present a paper, submit a two-page abstract no later than November 1, 1985. Individuals wishing to chair a session, or serve as a discussant should write indicating their arenas of interest and expertise to: Allan Young, Vice President, EFA Program, School of Management, Syracuse University, Syracuse, NY 13210.

Call for Papers: The International Conference on Marketing and Development will be held in Istanbul, Turkey, September 1–4, 1986. Topics will be Development and the Role of Marketing; Managerial Issues in Marketing and Development; and Consumer Issues in Development. Submit abstracts of 300–500 words no later than November 4, 1985 to the Conference Coordinator, Erdógan Kumcu, Department of Marketing, Ball State University, Muncie, IN 47306 (telephone 317+285–5186).

Call for Papers: The Centre HEC–ICA (CESA) is organizing a conference on Research in International Finance to be held June 19–20, 1986, in Jouy-en-Josas near Paris. Papers on all aspects of international finance are invited. Sessions are planned on international finance management, money, capital markets, and comparative national studies. A special session will be devoted to computer applications designed to assist international financial management. Papers, or detailed abstracts, should be submitted in duplicate before December 15, 1985 to Professors Bernard Marois and Bruno Solnik, Centre HEC–ISA, B.P. 100, 78350 Jouy en Josas, France (telephone:; (33-3) 956 8000).

The sixth International Meeting of the Association d'Econometrie Appliquee/Applied Econometrics Association will be held at Erasmus University, Rotterdam, December 16–17, 1985. For further information, contact M. Lamure, Université de Lyon 1, Bâtiment 101, 43 Bd du 11 Novembre 1918, 69622 Villeurbanne, Cedex, France.

The Third World Congress for Soviet and East European Studies, cosponsored by the American Association for the Advancement of Slavic Studies and the International Committee for Soviet and East European Studies, will be held at the Sheraton Washington Hotel, Washington, D.C., October 30–November 4, 1985. For information, contact AAASS, 128 Encina Commons, Stanford University, Stanford, CA 94305 (telephone 415+497-9668).

The Center for Great Plains Studies announces its tenth annual symposium, March 20–22, 1986. The conference, "The Meaning of the Plains Indian Past for Present Plains Culture," will focus on European transformations of social and symbolic forms in Plains Indian cultures; Indian and non-Indian cultural relationships; and the cultural and economic development of indigenous peoples in the face of Euro-American incursions into their territory and culture. For further information, contact Professor Paul Olson, Center for Great Plains Studies, University of Nebraska, 1213 Oldfather Hall, Lincoln, NB 68588–0314.

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to airfare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for application to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 228 East 45 Street, New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

Deaths

Kent T. Healy, professor emeritus of economics, Yale University, January 9, 1985.

Ralph C. Jones, professor emeritus of economics, Yale University, November 20, 1984.

Tjalling C. Koopmans, Alfred Cowles professor emeritus of economics, Yale University, February 26, 1985.

Elsie M. Knoer, assistant professor of economics, University of California-Davis, January 30, 1985.

Raymond Lubitz, assistant director, Federal Reserve, July 13, 1984.

Paul Rosenstein-Rodan, Boston, MA, April 28, 1985.

Retirements

Rudolph C. Blitz, professor of economics, Vanderbilt University, July 1985.

Jesse Burkhead, Maxwell professor of economics and public administration, Syracuse University, May 1985.

George H. Cleaver, vice president of research, Merrill Lynch, August 1984.

Bruce Glassburner, professor of economics, University of California-Davis, July 1, 1984.

Promotions

David Evans: associate professor of economics, Fordham University.

Asim Erdilek: professor of economics, Case Western Reserve University, July 1, 1985.

Steven D. Gold: director, Fiscal Affairs Program, National Conference of State Legislatures, March 1, 1985.

Cliff J. Huang: professor of economics, Vanderbilt University, September 1984.

Paul W. Kuznets: professor of economics, Indiana University-Bloomington, July 1, 1985.

Mark Machina: associate professor of economics, University of California-San Diego, July 1, 1984.

Rebecca A. Maynard: vice president, deputy director of research, Mathematica Policy Research, Inc., February 1, 1985.

Karl D. Meilke: professor of agricultural economics, University of Guelph, July 1985.

Frank J. Navratil: professor of economics, John Carroll University, September 1, 1984.

Lawrence Schwartz: technical manager, Mathtech, Division of Mathematica Consulting and Research, Arlington, VA, January 1985.

Joel Sobel: associate professor of economics, University of California-San Diego, July 1, 1984.

Halbert White: professor of economics, University of California-San Diego, July 1, 1984.

Arlington W. Williams: associate professor of economics, Indiana University-Bloomington, July 1, 1985.

Administrative Appointments

Theodore Groves: chair, economics department, University of California-San Diego, September 1, 1984.

John A. Sawyer: acting dean, faculty of management studies, University of Toronto, July 1, 1985–June 30, 1986.

New Appointments

Gerald Y. Agbegha: visiting assistant professor of economics, John Carroll University, January 20, 1985.

Ronald S. Blum, University of Wisconsin-Madison: visiting lecturer in economics, Indiana University-Bloomington, August 1985.

Harris Dellas: assistant professor of economics, Vanderbilt University, August 1985.

Edward A. Downe, St. John's University: visiting assistant professor of economics, John Carroll University, January 20, 1985.

Joseph Grunwald, The Brookings Institution: president, Institute of the Americas, August 1984.

Margaret E. Guerin-Calvert, Board of Governors of the Federal Reserve System: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, January 1985.

Masanori Hashimoto, University of Washington: professor of economics, Indiana University-Bloomington, August 1985.

Helen H. Jenson, University of Maryland: visiting assistant professor of economics, Iowa State University, Spring 1985.

David R. Johnson, Bank of Canada: assistant professor of economics, Wilfrid Laurier University, July 1, 1985.

Thomas Klitgaard: economist, Industrial Economies Division, International Research Department, Federal Reserve Bank of New York, October 3, 1984.

Geoffrey Newman, University of Toronto: associate professor of economics, Drew University, September 1, 1985.

S. Wayne Passmore: economist, Banking Studies Division, Banking Studies Department, Federal Reserve Bank of New York, August 29, 1984.

Charles A. Pigott: research officer and senior economist, Research and Statistics Function, Federal Reserve Bank of New York, March 1, 1985.

Kevin D. Salyer: instructor in economics, Vanderbilt University, August 1985.

Joel Scheraga, Rutgers University: visiting assistant professor of economics, Princeton University, September 1, 1985–July 1, 1986.

David Schirm, James Madison University: assistant professor of economics, John Carroll University, September 1, 1984.

Marilyn Skiles: economist, Developing Economies Division, External Financing Department, Federal Reserve Bank of New York, December 17, 1984.

Edward J. Sullivan: assistant professor of economics, Fordham University, September 1985.

Antipol Bhanich Supapol, University of Western Ontario: lecturer, department of economics, Wilfrid Laurier University, July 1, 1985.

Christopher J. Waller, Washington State University: lecturer in economics, Indiana University-Bloomington, August 1985.

Laurence Weiss: associate professor of economics, University of California-San Diego, July 1, 1985.

Timothy M. Weithers: assistant professor of economics, Fordham University, September 1985.

John D. Wilson, University of Wisconsin-Madison: associate professor of economics, Indiana University-Bloomington, August 1985.

Wing T. Woo: assistant professor of economics, University of California-Davis, July 1, 1985.

Thomas J. Zlatoper, West Virginia University: John Carroll University and research associate, Regional Economic Issues Program, Cleveland Federal Reserve Bank, September 1, 1984.

Leaves for Special Appointments

Roy J. Gardner, Indiana University: Alexander von Humboldt Fellow, University of Bonn, 1985-86.

Resignations

Peyton J.Marshall III, assistant professor of economics, Vanderbilt University, August 1985.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, use the following style:

A. Please use the following categories (please—do not send public relation releases):

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), new institution and the date at which the change will occur.

C. Type each item on a separate 3×5 card.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

All items and information should be sent to the Assistant Editor, *American Economic Review*, 169 Nassau Street, Princeton, NJ 08542–7067.

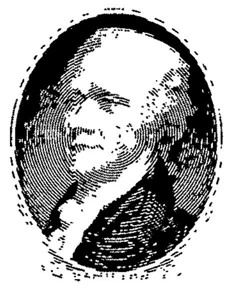
Please mention The American Economic Review *When Writing to Advertisers*

Economics of Worldwide Stagflation
Michael Bruno and Jeffrey Sachs
"This is *serious* supply-side economics . . . An excellent and important book."–Robert M. Solow
$25.00

Racial Conflict and Economic Development
W. Arthur Lewis
Lewis discerns the ways in which race and economics affect individuals and groups, bringing a personal viewpoint to the problems faced by both less-developed and more-developed countries.
The W. E. B. Du Bois Lectures $12.50

Tax Expenditures
Stanley S. Surrey and Paul R. McDaniel
"This book is comprehensive, challenging, and constructive, and it fills a big void in the tax literature. It will interest economists, political scientists, lawyers, accountants, and legislators."–Joseph Pechman
$27.50

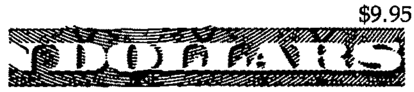
Housing and Neighborhood Dynamics
A Simulation Study
John F. Kain and William C. Apgar, Jr.
Using the Harvard Urban Development Simulation Model, the authors assess the effects of programs that provide direct assistance to low-income property owners in an attempt to arrest neighborhood decline and encourage revitalization.
Harvard Economic Studies, 157 $27.50



Family Firm to Modern Multinational
Norton Company, a New England Enterprise
Charles W. Cheape
The history of Norton Company, so carefully detailed here from company records and interviews, illustrates both the continuity and change important in the evolution of large-scale business in the United States.
Harvard Studies in Business History, 36 $25.00

Equality in America
The View from the Top
Sidney Verba and Gary Orren
Based on a study of leaders from all significant sectors of American society, this book examines American attitudes toward equality by placing those beliefs in historical context and demonstrating a relationship between political and economic equality.
$25.00 cloth; $12.50 paper

Tax Revolt
Something for Nothing in California
Enlarged Edition
David O. Sears and Jack Citrin
"This book is the first substantial study of the California tax revolt. It is well crafted and well written and has great depth and sophistication. The topic is of intense interest and importance . . . A significant work."–Walter Dean Burnham
$9.95 paper

A Treatise on the Family


Gary S. Becker
"This truly pathbreaking book marries techniques and problems hitherto regarded as utterly incompatible–rigorous economic reasoning to understanding the family. The marriage is astoundingly productive. It is destined to affect the foundations of every science dealing with human behavior."–Milton Friedman
$7.95 paper

Harvard University Press
79 Garden Street, Cambridge, Mass. 02138

The American Economic Association is pleased to announce its Minority Fellowship Awards for 1985-86:

Continuing Fellows:

Darrell Gaskin, Massachusetts Institute of Technology
Sam Green, Northwestern University

Newly-appointed Fellows:

Jaime Alvarez, University of Wisconsin at Madison
Vernita Fort, University of Maryland
Juan Gonzalez, Massachusetts Institute of Technology

The Fellowships are awarded to minority students who are enrolled in their first and second years of study in a Ph.D. program in Economics. They are funded by a grant from The Rockefeller Foundation and administered by the AEA Committee on the Status of Minority Groups in the Economics Profession (CSMGEP).

For more information, contact: Julie Marsh, Registrar, AEA Fellowship Program, Department of Economics, P.O. Box 1972 Yale Station, New Haven, CT 06520-1972. Phone: 203-436-3303.

Begin making plans to attend the

Annual Meeting of The American Economic Association, Centennial Celebration

(in Conjunction with Allied Social Science Associations)

to be held in

NEW YORK, NY

Dec. 28-30, 1985

The Employment Center opens Friday, December 27.

See the Notes section of the September *AER* for the American Economic Association's preliminary program.

The 1986 meeting will be held in New Orleans, LA, December 28-30.
